Making choices

Modelling the English dative alternation

Cover photos by Harm Lourenssen

ISBN 978-94-6191-275-6

Making choices

Modelling the English dative alternation

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. mr. S.C.J.J. Kortmann,
volgens besluit van het college van decanen,
in het openbaar te verdedigen op maandag 18 juni 2012
om 13.30 uur precies

door

Daphne Louise Theijssen

geboren op 25 juni 1984
te Uden

Promotor:
Prof. dr. L.W.J. Boves

Copromotoren:
Dr. B.J.M. van Halteren
Dr. N.H.J. Oostdijk

Manuscriptcommissie:
Prof. Dr. A.P.J. van den Bosch
Prof. Dr. R.H. Baayen (Eberhard Karls Universität Tübingen, Germany
& University of Alberta, Edmonton, Canada)
Dr. G. Bouma (Rijksuniversiteit Groningen)

Nobody said it was easy
It's such a shame for us to part
Nobody said it was easy
No one ever said it would be this hard
Oh take me back to the start

I was just guessing at numbers and figures
Pulling your puzzles apart
Questions of science; science and progress
Do not speak as loud as my heart

The Scientist – Coldplay

Dankwoord / Acknowledgements

Wanneer ik mensen vertel dat ik een proefschrift heb geschreven, kijken ze me vaak vol bewondering aan. Maar een proefschrift schrijf je niet alleen. Op allerlei gebieden ben ik geholpen tijdens de vier jaar waarin ik mijn promotieonderzoek heb gedaan. Een woord van dank is daarom op zijn plaats.

Allereerst wil ik mijn promotor, Lou Boves, bedanken. Lou is een geïnteresseerde, eerlijke en praktisch ingestelde begeleider. Mede dankzij hem is het gelukt mijn proefschrift binnen vier jaar af te ronden. Lou, ontzettend bedankt voor je eindeloze tijd voor doeltreffende reviews, en voor het aanbrengen van focus in mijn project.

Ten tweede zijn daar mijn co-promotoren, Hans van Halteren en Nelleke Oostdijk. Met Hans heb ik een aantal mastercursussen verzorgd, waarin ik ontdekte hoe goed hij de moeilijkste stof kan uitleggen. Tijdens onze samenwerking kwam hij regelmatig mijn kantoor binnenstormen, om pas na een uitgebreide samenvatting van zijn reis per openbaar vervoer ter zake te komen. Maar ik heb inhoudelijk veel van Hans geleerd. Met Nelleke heb ik wat minder inhoudelijk samengewerkt, maar ze was altijd bereid om mijn artikelen te reviewen, om mee te denken over de richting van het onderzoek, en om de taalkunde in beeld te houden wanneer Lou, Hans en ik ons verloren in de statistische modellen. Hans en Nelleke, bedankt voor de deur die altijd open stond.

Degene van wie ik in mijn dagelijks leven als promovenda het meest geleerd heb, is zonder twijfel mijn kantoorgenoot en paranimf Suzan Verberne. Tijdens mijn masteropleiding heb ik stage gelopen in Suzan's promotieproject, en toen ik aan mijn eigen promotieproject begon kreeg ik een werkplek bij Suzan op de kamer. Er volgden vier jaren met veel thee (maar liever geen kersen), veel 3FM (maar liever geen Anouk en Van Velzen) en veel gegiebel en gelach (maar liever niet als Lou net binnen kwam). Direct of indirect heb ik van Suzan leren programmeren, plannen, schrijven en samenwerken. Suzan, ik ga je echt missen. Ontzettend bedankt voor alles!

Mijn tweede paranimf, Noortje Groenen, heeft me vooral in het begin van mijn promotie veel bijgestaan met lunchdates in de Refter. Later volgde ook nog de nodige ontspanning tijdens de wintersport in Oostenrijk. Noor, bedankt voor je eeuwige interesse, en voor je hulp tijdens de voorbereiding van mijn verdediging.

Dat ik met zoveel plezier heb gewerkt aan mijn proefschrift is vooral ook te danken aan mijn collega's bij Taalwetenschap. De dagelijkse koffiepauzes, de

jaarlijkse afdelingsuitjes, en de vele fruitmomentjes braken de sleur. Ik wil in het bijzonder Eric Sanders, Maaike Jongenelen en Loes Oldenkamp bedanken voor de fijne gesprekken. Ook bedank ik graag Louis ten Bosch en Bert Cranen voor het meedenken over mijn onderzoek. En natuurlijk Christel Theunissen, Hella Cranen en Wies de Beijer van het secretariaat, die altijd voor de afdeling (en dus ook voor mij) klaar stonden.

In the Fall of 2010, I worked at the Linguistics Department at Stanford University in California, USA. I worked with one of the most inspiring professors I have ever met: Joan Bresnan. Thank you Joan, for being so open, passionate, funny and critical. I also enjoyed working with the other colleagues at the Spoken Syntax Lab, especially Jason Grafmiller, Rebecca Greene and Stephanie Shih. Furthermore, I thank Marilyn Ford from Griffith University, Australia, for our pleasant collaboration. I had a great time in California, mostly thanks to my lovely office mate Caroline Piercy. Spending time with her and Steve Lee, Chris Evenhuis and Penny Stolp really made me feel at home. Thanks, guys!

Ook buiten de universiteit heb ik veel steun gehad. Dankzij mijn vriendinnen van de studie Engels/Amerikanistiek en de meiden uit omgeving Uden heb ik ook een leven gehad buiten mijn promotieonderzoek. Bedankt Hilde, Julie (en Noor nogmaals), Margeertje, Kirsten, Marijke, Sabien, Merel, Janneke, Aafke, Janny en Isabelle, en de bijbehorende mannen, voor de afleiding tijdens de jaarlijkse weekendjes, de etentjes en de avondjes uit!

Aan het einde van dit dankwoord kom ik bij de mensen die het dichtst bij mij staan. Mijn ouders, mijn broer en schoonzus, en mijn schoonouders. Altijd hebben ze me gesteund in mijn keuzes en hiervoor ben ik ze ontzettend dankbaar. Dit geldt ook zeker voor mijn lieve vriend Harm, die de afgelopen vier jaar alle lusten en lasten met mij gedeeld heeft. Niet alleen in mijn promotieproject, maar ook in ons gezamenlijke project: de bouw van ons eigen huis. Op de momenten dat ik in het buitenland was vanwege conferenties, en toen ik in Amerika werkte, zorgde hij ervoor dat alles thuis gewoon doorging. Lieve Harm, dank je wel voor alles.

Contents

1

Introduction

Traditional linguistic theories have attempted to design deterministic rules that would account for all-and-only the sentences of a language that are deemed 'grammatical'. While acknowledging the fact that language use may be variable (graded), conventional theories assume that the underlying human grammar is categorical: a sentence is either grammatical, or it is not. The idea has now gained ground that 'grammaticality' is a graded concept itself, and that human language behaviour may be essentially probabilistic in nature. A probabilistic theory of language can take various forms (e.g. the studies presented in Bod, Hay, & Jannedy, 2003), and resembles memory-based and exemplar-based models of language (e.g. Daelemans & van den Bosch, 2005; Gahl & Yu, 2006).

It is not surprising that many linguists have now moved from studying the dichotomy ('grammatical' and 'ungrammatical'), to studying variation in language. One obvious example of variation is syntactic alternation, in which there are different grammatical constructions that could be used to express the same core semantics. The alternative grammatical constructions are competing, and language users choose (subconsciously) among these options. For instance, speakers of English can choose between the *s*-genitive, as in *John's dog*, and the *of*-genitive, in *the dog of John* (e.g. Rosenbach, 2003).

One of the best-studied syntactic alternations is the **dative alternation** in English, in which speakers and writers can choose between structures with a prepositional dative (example 1) or double object structure (example 2):

1. The evil queen gives the poisonous apple to Snow White.
2. The evil queen gives Snow White the poisonous apple.

The dative alternation is also known by many other names, for instance the 'diathesis alternation' and the 'ditransitive construction'. In this thesis, we use the term 'dative alternation'. The two objects of the verb will be referred to as

the **recipient** (*Snow White* in examples 1 and 2) and the **theme** (*the poisonous apple* in the examples).

There are two additional options in the alternation: the reversed prepositional dative construction (e.g. *I gave to him a book*) and the reversed double object construction (e.g. *I gave it him*). Also, the alternation can occur with prepositions other than *to*, e.g. with *for* (the benefactive alternation, cf. Theijssen et al., 2009). These constructions are fairly infrequent, especially compared to the two constructions in examples 1 and 2. In order to prevent data sparseness problems (as for instance those in Theijssen et al., 2009), we therefore limit ourselves to the alternation between the two most frequent options, being the alternation between the double object construction and the prepositional dative construction with ***to***, both in the default object ordering (examples 1 and 2). All mentions of 'prepositional dative' in this thesis thus refer to the variant with *to* only, unless explicitly indicated otherwise.

Parallels to the dative alternation occur in various languages other than English, for example in Dutch (e.g. Colleman, 2006), Greek (e.g. Anagnostopoulou, 2005), Spanish (e.g. Beavers & Nishida, 2010) and Brazilian Portuguese (e.g. Gomes, 2003). In this thesis, we take the dative alternation in English as a case study, focussing mostly on British English. The set of remaining dative constructions also contains instances with a clausal object (e.g. *tell him how nice he is*), with a phrasal verb (e.g. *to hand over*), in passive voice (e.g. *He was given a book*), in imperative clauses (e.g. *Give him the book!*), and in interrogative clauses (e.g. *Did he give you the book?*). These special cases may be influenced by syntactic variation other than the dative alternation, such as passive versus active voice, declarative versus interrogative mode and the placement of particles. One way to take this into account is to control for any other syntactic variation when carrying out the statistical analyses. However, the default syntactic structure is the most frequent by far, which would lead to serious imbalance in the data. For this reason, we follow Bresnan, Cueni, Nikitina, and Baayen (2007) and exclude all instances with the aforementioned characteristics.

The dative alternation has been the object of study in several subdisciplines of linguistics, e.g. for corpus linguistics (e.g. Bresnan et al., 2007), psycholinguistics (e.g. Bresnan & Ford, 2010), first language acquisition (e.g. de Marneffe, Grimm, Arnon, Kirby, & Bresnan, 2012), second language acquisition (e.g. Babanoğlu, 2007), sociolinguistics (e.g. Szmrecsányi, 2010) and historical linguistics (e.g. Wolk, Bresnan, Rosenbach, & Szmrecsányi, 2012). Previous research has already found sets of predictive syntactic, semantic, and discourse-related features that appear to influence the likelihood of the two dative constructions. In general, speakers and writers show a tendency to place animate nouns before inanimate nouns, shorter constituents before longer ones,

discourse given before discourse new, pronouns before nonpronouns and definite before indefinite. These features are introduced in more detail in Section 1.1.

For the last decade, many researchers in the various subdisciplines have started using multivariate models to study the role of the features suggested in the literature (e.g. Arnold, Wasow, Losongco, & Ginstrom, 2000; Bresnan et al., 2007). Such models, usually (logistic) regression models, allow linguists to study the relevance of the features in one integrated model, instead of studying the influence of the features in isolation. Although the use of these advanced statistical techniques has led to interesting insights in research on syntactic alternations, it also led to some complications. Across the subdisciplines, linguists studying syntactic alternation have to make choices: which features to include in the study (*variable selection*), how to define and annotate the features used (*feature definition*), how to obtain an annotated data set that is sufficiently large (*data collection*), how to study the alternation across different speaker groups (*comparison of speaker groups*) and how to interpret the models found with various techniques (*model interpretation*). In this thesis, we address the various methodological choices that linguists can make when studying the dative alternation. The research is interdisciplinary: It involves corpus linguistics, psycholinguistics and sociolinguistics.

The remainder of this chapter contains a brief overview of previous research, an outline of the research questions treated in the subsequent chapters of this thesis, and a summary of the data sets used.

1.1 Previous research

There is already a vast body of research on the English dative alternation. This section presents the studies from which we adopted and adapted the features used in this thesis. Also, we list some literature on the role that the verb (e.g. *give* in Examples 1 and 2) plays in the dative alternation. Existing work that is related to the individual research questions (introduced in Section 1.2) is provided in the Related Work sections of the subsequent chapters.

Features

Many researchers have tried to find features with which to explain the alternation between the two dative constructions. Some have argued that a change in syntactic structure is likely to cause a change in meaning (Bolinger, 1977; Pinker, 1989; Levin, 1993). However, empirical studies have shown that in spontaneous speech, speakers sometimes employ both alternatives in the same context, using the same words, but with the other syntactic construction

(Davidse, 1996; Bresnan & Nikitina, 2009). Research has also been directed at other semantic features that may affect the likelihood of the two constructions. Quirk, Greenbaum, Leech, and Svartvik (1972, p. 843) for instance mentioned that 'indirect objects are typically animate'. Collins (1995, p. 47–48) found a discourse effect, seeing a 'strong likelihood' that the recipient will be 'informationally given' and the theme 'informationally new' in the double object construction.

Bresnan et al. (2007) combined the features suggested in individual studies in a multivariate model; they found that, everything else being equal:

> animate objects are usually mentioned before inanimate objects,
> definite objects usually before indefinite objects,
> given objects usually before nongiven objects
> shorter objects usually before longer objects,
> and pronouns usually before nonpronouns

These features, and the probabilistic framework in which they are usually exploited, do not necessarily reflect the processes that take place in our brains, but they are able to explain a lot of the variance in a data set: Bresnan et al. (2007) fitted various logistic regression models for the dative alternation based on 2360 instances they extracted from the three-million word Switchboard Corpus of transcribed American English telephone dialogues (Godfrey, Holliman, & McDaniel, 1992), and were able to predict 94% of the choices in an unseen part of the corpus using a limited set of features. The prediction accuracy of their model was significantly better than the majority baseline of always selecting the double object construction: 79% (1859 of 2360 instances in the Switchboard corpus). Also, the relevance of the features employed in Bresnan et al. (2007) was already established independently in psycholinguistic research (e.g. Bock & Irwin, 1980; Bock, Loebell, & Morey, 1992; Prat-Sala & Branigan, 2000), in other corpus studies (e.g. Weiner & Labov, 1983; Givón, 1984; Estival & Myhill, 1988; Thompson, 1990, 1995; Collins, 1995; Snyder, 2003; Gries, 2003; Szmrecsányi, 2005, 2006), and in studies that combine experimental and corpus data (e.g. Arnold et al., 2000; Rosenbach, 2003, 2005).

We adapted the set of features in Bresnan et al. (2007), as summarised in Table 1.1. Not all possible object–feature combinations are included in the table (and in our research) because the bias in a combination can be too strong to keep it in regression models: most themes are inanimate and nonlocal (3rd person), and most recipients are concrete. The length factor (shorter precedes longer) is a proxy for syntactic complexity or *end weight* (Behaghel, 1909).

Table 1.1: Features used in this thesis, adapted from Bresnan et al. (2007).

Feature	Values	Definition
Animacy of the recipient	animate, inanimate	human or animal, or not
Concreteness of theme	concrete, inconcrete	prototypically 'concrete', or not
Definiteness of recipient	definite, indefinite	definite pronoun, proper name or noun preceded by definite determiner, or not
Definiteness of theme	definite, indefinite	Id.
Discourse givenness of recipient	given, nongiven	mentioned/evoked ≤20 clauses before, or not
Discourse givenness of theme	given, nongiven	Id.
Length difference (log of ratio)	*interval*	ln(# words theme) - ln(#words recipient)
Number of recipient	singular, plural	singular in number, or plural
Number of theme	singular, plural	Id.
Person of recipient	local (1st/2nd), nonlocal (3rd)	first or second person (*I*, *you*), or not
Pronominality of recipient	pronoun, nonpronoun	headed by a pronoun, or not
Pronominality of theme	pronoun, nonpronoun	Id.
Semantic class	transfer of possession, future transfer of possession, prevention of possession, communication, abstract	*give it some thought* is abstract, *tell him a story* is communication *give*/*deny*/*promiss him the book* is transfer
Structural parallelism in dialogue	yes, no	preceding instance is prepositional dative, or not

Verb

It is generally known that many ditransitive verbs have a strong preference for one of the two constructions. For instance, Gries and Stefanowitsch (2004) investigated the effect of the verb in 1772 instances from the ICE-GB Corpus (Greenbaum, 1996). When predicting the preferred dative construction for each verb, 82.2% of the instances could be assigned the correct construction. Using verb bias as a predictor outperforms the majority baseline of 65.0% in this ICE-GB data set. Gries and Stefanowitsch (2004, p. 104) suggested that the double object construction 'should prefer verbs of direct face-to-face transfer, while the *to*-dative should prefer verbs of transfer over distance'.

Bresnan et al. (2007) accounted for the effect of verb preferences with the help of a mixed-effect logistic regression model (or *mixed model*), including verb sense as a random effect. They defined the verb sense as the verb lemma together with its semantic verb class. As mentioned in the previous section on Features, their model was able to predict correctly 94% of previously unseen data in the Switchboard corpus.

Seeing the influence that the verb has on the choice for one of the two dative constructions, the models presented in this thesis also take into account the role of the verb.[1] In most cases, we include the verb sense (or the verb) as a random effect in a mixed-effect logistic regression model, following the approach in Bresnan et al. (2007).

1.2 Research questions and thesis outline

As mentioned in the beginning of this chapter, linguists studying the dative alternation can make choices with respect to variable selection, feature definition, (automatic) data collection, comparison of speaker groups and model interpretation. These five aspects are the topics of our five research questions, each treated in a separate chapter.

Variable selection

In Chapter 2, we aim at answering the question:

Is it justified to report only one 'optimal' regression model, if models can be built in several different ways?

We address this question by building regular and mixed (i.e. containing a random effect) logistic regression models in order to explain the British English dative alternation in corpus data. The models were constructed with three

[1] The only exception is the research in Chapter 3, which focusses on the definition of a different feature: concreteness.

different variable selection approaches. In total, we thus build six logistic regression models for the same data set. We compare the models with respect to the regression coefficients found for the features.

To establish the quality of the six regression models, they are used to predict the (log) odds that a given data instance is prepositional dative, not double object. We evaluate the models found with respect to their prediction accuracy and the concordance C. For the prediction accuracy, we cut off the odds at 0: all instances with odds > 0 are classified as prepositional dative, all others as double object. The prediction accuracy is the proportion of instances that is classified correctly. The concordance C is less crude: it is the proportion of all possible pairs of a prepositional dative and a double object instance, for which the regression model indeed assigns the highest odds to the prepositional dative.

Feature definition: concreteness

During the manual annotation of the data in Chapter 2, we discovered that concreteness is one of the most difficult features to annotate, mostly because it is so hard to establish a clear definition. The κ scores in Theijssen et al. (2009) and Chapter 4 show that the inter-annotator agreement for concreteness is indeed one of the lowest (together with discourse givenness). In Chapter 3, we therefore investigate the effect of using different definitions for concreteness, addressing the research question:

What is the impact of different instantiations of the definition of the feature 'concreteness' on the actual labels given to corpus data, and on the outcome of syntactic research using this data?

We compare different definitions of concreteness, and use them in different implementations to annotate nouns in two corpus data sets. One data set is used for an intrinsic evaluation, in which we compare the actual concreteness values assigned by the different approaches. The other data set is employed for an extrinsic evaluation, in which we use the concreteness values in regression models of the dative alternation. The various approaches to concreteness differ in the definition used, in the measurement scale of the values that can be assigned (interval, ordinal, nominal), the noun level they take as basis (token, sense or type) and the manner in which the values are assigned (manually, automatically, or semi-automatically). The chapter also contains a crowdsourcing study to investigate how (non-linguist) humans rate the concreteness of nouns.

Automatic data collection

In Chapter 4, we present and evaluate an approach for automatically obtaining a data set for studying the dative alternation. Our research question is:

Is data that is obtained and annotated automatically suitable for linguistic research, even if the data may contain a certain proportion of errors?

We automatically create two richly annotated data sets for studying the English dative alternation, making use of existing corpora. The two data sets are evaluated intrinsically and extrinsically. In the intrinsic evaluation, we compare the data sets to gold standard data. First, we establish the precision, recall and F-score of our approach to finding dative candidates automatically, and second, we find the accuracy and κ-score of our automatic feature extraction. The extrinsic evaluation consists of building logistic regression models with the two data sets, and comparing them.

Comparison of speaker groups

Chapter 5 contains a linguistic study that combines aspects of corpus linguistics, psycholinguistics and sociolinguistics. We compare the dative alternation in different speaker groups, aiming to answer the research question:

What are the differences and/or similarities in the dative alternation of British, American and Australian language users varying in age and gender?

This chapter presents a corpus and a judgement study of the dative alternation, in a framework in which we assume that syntactic structure is influenced by linguistic factors of which the relative importance may vary across different speaker groups, thus introducing extra-linguistic factors. With regression models, we compare the dative alternation of British, American, and Australian speakers of English varying in age and gender.

Model interpretation

In Chapter 6, we focus on the following research question:

How suitable are regression models, Bayesian networks and memory-based learning for studying the dative alternation?

We use three different approaches to model the dative alternation in a large corpus data set. The suitability of the approaches is tested against three criteria: the quality of the model in terms of classification accuracy, the interpretability of the model in linguistic research, and the actual classification of individual cases by the model.

Chapter 7 contains a summary of the core chapters, followed by our general conclusion and suggestions for future research.

1.3 Overview of data sets used

The research in this thesis is based on various data sets. This section presents an overview of the data sets used, the type of data they contain and their relation to each other. Most data sets were extracted from existing corpora, some were collected through questionnaires. The data matrices (the instances with their feature values) that we collected ourselves will be made available on `http://daphnetheijssen.ruhosting.nl/downloads`. For the corpus data sets, the instances include a pointer to the location in the corpus from which it was extracted; licence agreements prevent making the full corpora available.

Corpus data sets found using manual procedures

Traditionally, data is extracted from existing corpora by looking up instances of interest either completely manually, or making use of the annotations available in the corpora, being manually added or checked. A number of such data sets are used in this thesis:

- In Chapter 2, we introduce our Ice-Trad data set, consisting of 930 dative instances from the one-million-word (manually annotated) British component of the ICE Corpus, the ICE-GB (Greenbaum, 1996).
- Chapter 3 uses a subset of Ice-Trad, consisting of the 619 instances in Ice-Trad that could be parsed automatically with the approach described in Chapter 4: the Dative data set.
- Chapter 3 also uses a data set of the 68,484 nouns annotated with a WordNet word sense in the (manually annotated) SemCor corpus (Miller, Leacock, Tengi, & Bunker, 1993): the SemCor data set.
- In Chapter 4, we use Ice-Trad again, together with Swb-Trad, the 2,349 datives collected from the a Switchboard corpus of spoken telephone dialogues in American English (Godfrey et al., 1992), being a corrected version of the original set described in Bresnan et al. (2007).[2]
- In Chapter 5, we use a combination of Ice-Trad and Swb-Trad, containing the 2,541 instances in spontaneous speech with a limited set of verbs shared by both sets.

[2] I thank Prof. Joan Bresnan for sharing this data set.

Corpus data sets semi-automatically found

In Chapter 4, we introduce a method for automatically obtaining data to study the dative alternation. This leads to the semi-automatic data sets, which are instances that were automatically found, manually checked, and automatically annotated for the features:

- We obtained 633 dative instances from the ICE-GB corpus (ICE-SEMI) in Chapter 4.
- Also in Chapter 4, we obtained 1,292 dative instances from the Switchboard corpus (SWB-SEMI).
- In Chapter 6, we applied the approach to the 100-million-word British National Corpus (BNC Consortium, 2007), leading to a set of 11,784 dative instances.

Corpus data sets automatically found

Chapter 4 introduces two data sets that were obtained completely automatically:

- ICE-AUTO, containing 889 dative candidates found in the ICE-GB corpus.
- SWB-AUTO, containing 2,694 dative candidates in the Switchboard corpus.

Rating data from questionnaires

Besides corpus data sets, this thesis also includes two data sets obtained through questionnaires:

- Workers on the crowdsourcing platform Amazon Mechanical Turk[3] rated the concreteness of 1,600 nouns in Chapter 3.
- In Chapter 5, 3,450 ratings of dative sentences in context were collected through a web-based questionnaire.

[3] `http://www.mturk.com`

2

Variable selection

Edited from: Theijssen, D. (2010). Variable selection in Logistic Regression: The British English dative alternation. In T. Icard & R. Muskens (Eds.), *Interfaces: Explorations in Logic, Language and Computation* (Vol. 6211 of Springer Lecture Notes in Artificial Intelligence, pp. 87–101). ISBN: 978-3-642-14728-9.

Abstract

This chapter addresses the problem of selecting the 'optimal' variable subset in a logistic regression model for a medium-sized data set. As a case study, we take the British English dative alternation. With 29 explanatory variables taken from the literature, we build two types of models: one with the verb sense included as a random effect, and one without a random effect. For each type, we build three different models by including all variables and keeping the significant ones, by successively adding the most predictive variable (forward selection), and by successively removing the least predictive variable (backward elimination). Seeing that the six approaches lead to six different variable selections (and thus six different models), we conclude that the selection of the 'best' model requires a substantial amount of expertise in linguistics and statistics.

2.1 Introduction

There are many linguistic phenomena that researchers have tried to explain on the basis of features on several different levels of description (semantic, syntactic, lexical, etc.), and it can be argued that no single level can account for all observations. Probabilistic modelling techniques can help in combining these partially explanatory features and testing the combination on corpus data. A popular – and rather successful – technique for this purpose is logistic regression modelling. However, *how* exactly the technique is best employed for this type of research remains an open question.

Statistical models built using corpus data do precisely what they are designed to do: find the 'best possible' model for a specific data set given a specific set of explanatory features. The issue that probabilistic techniques model data (while one would actually want to model underlying processes) is only aggravated by the fact that the variables are usually not mutually independent. As a consequence, one set of data and explanatory features can result in different models, depending on the details of the model building process.

Building a regression model consists of three main steps: (1) deciding which of the available explanatory features should actually be included as variables in the model, (2) establishing the coefficients (weights) for the variables, and (3) evaluating the model. The first step is generally referred to as *variable selection* and is the topic of the current chapter. Steps (1) and (3) are clearly intimately related.

Researchers have employed at least three different approaches to variable selection: (1) first building a model on all available explanatory features and then keeping/reporting those that have a significant contribution (e.g. Bresnan et al., 2007), (2) successively adding the most explanatory feature (forward), until no significant gain in model accuracy[1] is obtained anymore (e.g. Grondelaers & Speelman, 2007), and (3) starting with a model containing all available features, and (backward) successively removing those that yield the smallest contribution, as long as the accuracy of the model is not significantly reduced (e.g. Blackwell, 2005). In general, researchers report on only one (optimal) model without giving clear motivations for their choice of the procedure used.

In this chapter, we compare the three approaches in a case study: we apply them to a set of 930 instances of the British English dative alternation, taken from the British component of the ICE Corpus. The explanatory features (explanations suggested in the literature) are taken from Bresnan et al.'s work on the dative alternation in American English (Bresnan et al., 2007), as introduced in Chapter 1.

[1] Obviously, the accuracy measure chosen may have considerable impact on the result, but investigating this effect is beyond the scope of this Chapter and of this thesis.

Previous research (e.g. Gries & Stefanowitsch, 2004; Bresnan et al., 2007) has indicated that the verb or verb sense often predicts a preference for one of the two constructions. However, contrary to the fourteen explanatory features suggested by Bresnan et al., which can be treated as fixed variables because of their small number of values (often only two), *verb sense* has so many different values that it cannot be treated as a fixed variable in a regression model. Recently developed logistic regression models can handle variables with too many values by treating these as random effects (cf. West, Welch, & Galecki, 2007).[2] In order to examine the effect of building such *mixed models*, we create models with and without a random effect in each of the three approaches to variable selection described above. This leads to a total of six different models.

Our goal is to investigate whether it is justified to report only one 'optimal' regression model, if models can be built in several different ways. We will also pay attention to the role of a random effect in a model of syntactic variation built with a medium-sized set of observations. The case of the British English dative alternation is used to illustrate the issues and results.

The structure of this chapter is as follows: A short overview of the related work can be found in Section 2.2. The data is described in Section 2.3. In Section 2.4, we explain the method applied. The results are shown and discussed in Section 2.5. In the final Section (2.6), we present our conclusions.

2.2 Related work

Variable selection in building logistic regression models is an important issue, for which no hard and fast solution is available. In chapter 5 of Izenman (2008) it is explained that variable selection is often needed to arrive at a model that reaches an acceptable prediction accuracy and is still interpretable in terms of some theory about the role of the independent variables. Keeping too many variables may lead to overfitting, while a simpler model may suffer from underfitting. The risk of applying variable selection is that one optimizes the model for a particular data set. Using a slightly different data set may result in a very different variable subset.

Previous studies aimed at creating logistic regression models to explain linguistic phenomena have used various approaches to variable selection. For instance, Grondelaers and Speelman (2007) successively added the most predictive variables to an empty model, while Blackwell (2005) successively eliminated the least predictive variables from the full model. The main criticisms

[2] Another solution could be to convert the multinomial feature Verb into a numerical feature representing the bias of the verb towards one of the two constructions (cf. Gries & Stefanowitsch, 2004). However, this would require updating the feature values after seeing new data instances, while there is no need for such separate updating in a mixed model.

of these methods are (1) that the results are difficult to interpret when the variables are highly correlated, (2) that deciding which variable to remove or add is not trivial, (3) that all methods may result in different models that may be sub-optimal in some sense, and (4) that each provides a single model, while there may be more than one 'optimal' subset (Izenman, 2008).

A third approach to variable selection used in linguistic research is keeping only the significant variables in a complete model (cf. Bresnan et al., 2007). This is also what Sheather suggests in Sheather (2009, chapter 8). Before building a model, however, he studies plots of the variables to select those that he expects to contribute to the model. Where beneficial, he transforms the variables to give them more predictive power (e.g. by taking their log). After these preprocessing steps he builds a model containing all the selected variables, removes the insignificant ones, and then builds a new model. As indicated by Izenman (2008), variable selection on the basis of a data set may lead to a model that is specific for that particular set. Since we want to be able to compare our models to those found by Bresnan et al. (2007), who did not employ such transformations, we refrain from such preprocessing and we set out using the same set of variables they used in the variable selection process.

Yet another approach mentioned in Izenman (2008) is to build all models with each possible subset and select those with the best trade-off between accuracy, generalisability and interpretability. An important objection to this approach is that it is computationally expensive to carry out, and that decisions about interpretability may suffer from theoretical prejudice. For these reasons, we do not employ this method.

2.3 Data

Despite the fact that a number of researchers have studied the dative alternation in English (see Chapter 1), none of the larger data sets used is available in such a form that it enables the research in this chapter.[3] We therefore established our own set of instances of the dative alternation in British English. Since we study a syntactic phenomenon, it is convenient to employ a corpus with detailed (manually checked) syntactic annotations. We selected the one-million-word British component of the ICE Corpus, the ICE-GB, containing both written and (transcribed) spoken language (Greenbaum, 1996).

We used a Perl script to automatically extract potentially relevant clauses

[3] Although most of the data set used in Bresnan et al. (2007) is available through the `R` package `LanguageR`, the original sentences and some annotations are not publicly available because they are taken from an unpublished, corrected version of the Switchboard Corpus. At the time of writing and publishing this chapter, we had no access to the full data set; later, when writing Chapters 4 and 5, we did.

from the ICE-GB. These were clauses with an indirect and a direct object (double object) and clauses with a direct object and a prepositional phrase with the preposition *to* (prepositional dative). Next, we manually checked the extracted sets of clauses and removed irrelevant clauses such as those where the preposition *to* had a locative function (as, for example, in *Fold the short edges to the centre*.). As mentioned in Chapter 1, we also removed constructions with a preposition other than *to*, with a clausal object, in passive voice, with the reversed order (e.g. *She gave it me*), in an imperative or interrogative clause, and/or with a phrasal verb (e.g. *to hand over*). Coordinated verbs or verb phrases were also removed. The characteristics of the resulting data sets can be found in Table 2.1.

Table 2.1: Characteristics of the 930 instances taken from the ICE-GB Corpus

Medium	Double object	Prep. dative	Total
Spoken British English	406	152	558
Written British English	266	106	372
Total	672	258	930

2.4 Method

2.4.1 Explanatory features

We adopt the explanatory features and their definitions from Bresnan et al. (2007), and manually annotate our data set following an annotation manual based on these definitions (see Appendix). The features were already introduced in Table 1.1 in Chapter 1, but are repeated in Table 2.2 for the reader's convenience. Most features describe characteristics of the theme (*the poisonous apple* in Chapter 1) and the recipient (*Snow White*).

Our set includes one feature that was not used in Bresnan et al. (2007): *medium*, which tells us whether the construction was found in written or spoken text. It may well be that certain variables only play a role in one of the two media. In order to test this, we include the 14 (two-way) interactions between the features taken from Bresnan et al. and the medium.[4] Together with the feature *medium* itself, this yields a total number of 29 features.

[4]We are aware of the fact that there are other ways to incorporate the medium in the regression models, for instance by building separate models for the written and the spoken data. Since the focus of this chapter is on the three approaches in combination with the presence or absence of a random effect, we will limit ourselves to the method described.

Table 2.2: Explanatory features (th=theme, rec=recipient). All nominal explanatory features are transformed into binary variables with values 0 and 1.

Feature	Values	Description
1. rec = animate	1, 0	human or animal, or not
2. th = concrete	1, 0	with fixed form and/or space, or not
3. rec = definite	1, 0	definite pronoun, proper name or noun preceded by definite determiner, or not
4. th = definite	1, 0	Id.
5. rec = given	1, 0	mentioned ≤20 clauses before, or not
6. th = given	1, 0	Id.
7. length difference	-3.4-4.2	ln(#words in th) − ln(#words in rec)
8. rec = plural	1, 0	plural in number, or not (singular)
9. th = plural	1, 0	Id.
10. rec = local	1, 0	first or second person (*I*, *you*), or not
11. rec = pronominal	1, 0	headed by a pronoun, or not
12. th = pronominal	1, 0	Id.
13. verb = abstract	1, 0	*give it some thought* is abstract,
verb = communication	1, 0	*tell him a story* is communication,
verb = transfer	1, 0	*give him the book* is transfer
14. structural parallellism	1, 0	preceding instance prep. dative, or not
15. medium = written	1, 0	type of data is written, or not (spoken)

As mentioned in Section 2.1, we will build models with and without including *verb sense* as a random effect. Following Bresnan et al. (2007), we define the verb sense as the lemma of the verb together with its semantic class, e.g. *pay_a* for *pay* with an abstract meaning (*pay attention*) and *pay_t* when *pay* is used to describe a transfer of possession (*pay $10*). In total, our data set contains 94 different verb senses (derived from 65 different verbs). The distribution of the verb senses with 5 or more occurrences can be found in Table 2.3.

As predicted by Gries and Stefanowitsch (2004), many verbs show a bias towards one of the two constructions. The verb *give*, for instance, shows a bias for the double object construction, and *sell* for the prepositional dative construction. Only for *pay* and *send*, the bias differs for the different senses. For example, *pay* shows a clear bias towards the prepositional dative construction when it has an abstract meaning, but no bias when transfer of possession is meant. Nevertheless, we follow the approach in Bresnan et al. (2007) by taking the verb sense, not the verb, as the random effect.

Table 2.3: Distribution of verb senses with 5 or more occurrences in the data set. The verb senses in the right-most list have a clear bias towards the double object (DO) construction, those in the left-most for the prepositional dative (PD) construction, and those in the middle show no clear preference. The *a* represents *abstract*, *c communication* and *t transfer of possession*.

# DO > # PD			# DO ≈ # PD			# DO < # PD		
verb sense	DO	PD	verb sense	DO	PD	verb sense	DO	PD
give_a	255	32	*do_a*	8	10	*pay_a*	2	12
give_t	56	21	*send_c*	9	7	*cause_a*	5	8
give_c	66	10	*lend_t*	8	7	*sell_t*	0	10
tell_c	67	1	*pay_t*	6	5	*owe_a*	2	6
send_t	42	16	*leave_a*	5	4	*explain_c*	0	6
show_c	37	9	*write_c*	4	5	*present_c*	0	6
offer_a	24	9	*bring_t*	3	2	*read_c*	1	4
show_a	6	1	*hand_t*	3	2			
offer_t	6	0						
tell_a	6	0						
wish_c	6	0						
bring_a	4	1						

2.4.2 Variable selection

Using the values of the 29 explanatory features (fixed effect factors), we establish a regression function that predicts the natural logarithm (ln) of the odds that the construction c in clause j is a prepositional dative. The prepositional dative is regarded a 'success' (with value 1), while the double object construction is considered a 'failure' (0). The regression function for the models without a random effect is: (2.1):

$$\ln odds(c_j = 1) = \alpha + \sum_{k=1}^{29} (\beta_k V_{jk}) \ . \tag{2.1}$$

The α is the intercept of the function. $\beta_k V_{jk}$ are the weights β and values V_j of the 29 variables k. For the model with the random effect (for verb sense i), the regression function is:

$$\ln odds(c_{ij} = 1) = \alpha + \sum_{k=1}^{29} (\beta_k V_{jk}) + e_{ij} + r_i \ . \tag{2.2}$$

The random effect r_i is normally distributed with mean zero ($r_i \sim N(0, \sigma_r^2)$), independent of the normally distributed error term e_{ij} ($e_{ij} \sim N(0, \sigma_e^2)$). The optimal values for the function parameters α, β_k and (for models with a random

effect) r_i and e_{ij} are found with the help of Maximum Likelihood Estimation.[5]

The outcome of the regression enables us to use the model as a classifier: all cases with $\ln odds(c_j = 1) \geq t$ (for the models without a random effect) or $\ln odds(c_{ij} = 1) \geq t$ (for models with a random effect) are classified as prepositional dative, all with $\ln odds(c_j = 1) < t$ or $\ln odds(c_{ij} = 1) < t$ as double object, with t the decision threshold, which we set to 0. With this threshold, all instances for which the regression function outputs a negative ln odds are classified as double object constructions, all other instances as prepositional dative.

In the first approach, we include all 29 features in the model formula. We then remove all variables V_k that do not have a significant effect in the model output,[6] and build a model with the remaining (significant) variables.

For the second approach, being forward selection, we start with an empty model and successively add the variable that is most predictive. As Izenman (2008) explains, there are several possible criteria for deciding which variable to enter. We decide to enter the variable that yields the highest concordance C, which gives the proportion of the pairs with a positive (prepositional dative) and a negative (double object) instance, for which the regression function outputs a higher log odds for the positive instance than for the negative instance.[7] The concordance C is thus an evaluation measure for the quality of a model. We add the next most predictive variable to the model as long as it gives an improvement over the concordance C of the model without the variable. An interaction of variable V_k with *medium* is only included when the resulting concordance C is higher than the value reached after adding the main variable V_k.[8] Two concordance C values are considered different when the difference is higher than a threshold. We set the threshold to 0.002.[9]

For the third approach (backward elimination), we use the opposite procedure: we start with the full model, containing all 29 variables, and successively leave out the variable V_k that, after removal, yields the model with the highest concordance C that is not lower than the concordance for the model with V_k. When the concordance of a model without variable V_k does not differ from the concordance of the model without the interaction of V_k with *medium*, we remove the interaction. Again, concordance C values are only considered different when the difference exceeds a threshold (again set to 0.002).

We evaluate the models with and without random effects by establishing the

[5] We use the functions `glm()` and `lmer()` (Bates, 2005) in `R` (R Development Core Team, 2008).

[6] We use the p-values as provided by `glm()` and `lmer()`.

[7] We use the function `somers2()` created in `R` (R Development Core Team, 2008) by Frank Harrell.

[8] When including an interaction but not the main variables in it, the interaction will also partly explain variation that is caused by the main variables (Rietveld & van Hout, 2008).

[9] The threshold value has been established experimentally.

model fit (training and testing on all 930 cases) by calculating the percentage of correctly classified instances (accuracy) and the concordance (C). Also, we determine the prediction accuracy reached in 10-fold cross-validation (10 sessions of training on 90% of the randomised data and testing on the remaining 10%) in order to establish how well a model generalises to previously unseen data. In the 10-fold cross-validation setting, we provide the algorithms with the variables selected in the models trained on all 930 cases. The regression coefficients for these subsets of variables are then estimated for each separate training set.

The coefficients in the regression models help us understand which variables play what role in the dative alternation. We will therefore compare the coefficients of the significant effects in the models built on all 930 instances.

2.5 Results

2.5.1 Mixed models

Table 2.4 gives the model fit and prediction accuracy for the different regression models we built, including verb sense as a random effect. The prediction accuracy (the percentage of correctly classified cases) is significantly higher than the majority baseline (always selecting the double object construction) in all settings, also when testing on new data ($p < 0.001$ for the three models, Wilcoxon paired signed rank test).

Table 2.4: Number of variables selected, baseline accuracy, concordance C, and accuracies ± their confidence intervals (for model fit) or two times the standard deviations (for 10-fold cross-validation) for the regression models with *verb sense* as a random effect

			model fit (train=test)		*10-fold cv*
selection	#variables	baseline	C	accuracy	av accuracy
1. significant	6	0.723	0.979	0.935 (±0.016)	0.902 (±0.079)
2. forward	4	0.723	0.979	0.932 (±0.016)	0.890 (±0.069)
3. backward	4	0.72	0.978	0.928 (±0.017)	0.890 (±0.073)

When training and testing on all 930 instances, the mixed models reach high concordance and prediction accuracy (model quality). Their decrease in accuracy in the 10-fold cross-validation setting is (based on the confidence intervals and doubled standard deviations) not significant.[10]

[10] In the published article on which this chapter was based (Theijssen, 2010), there was a sig-

The significant effects for the variables selected in the three approaches are presented in Table 2.5. The directions of the main effects are the same as the results presented in (Bresnan et al., 2007) for American English.

Table 2.5: Coefficients of significant effects in (mixed) regression models with verb sense as random effect, trained on all 930 instances, *** p<0.001 ** p<0.01 * p<0.05. The (negative) effects above the horizontal line draw towards the double object construction, and the (positive) effects below it toward the prepositional dative construction.

Effect	1. significant		2. forward		3. backward	
length difference	-2.50	***	-2.44	***	-2.39	***
rec=animate	-1.01	*				
rec=given					-1.44	***
rec=given, medium=spoken			-0.94	*		
rec=given, medium=written			-1.74	***		
rec=local	-2.53	***	-1.82	***	-1.78	***
th=pronominal, medium=written	-1.79	*				
(intercept)	2.05	***	2.32	***	2.38	***
th=definite	1.78	***				
th=given			2.34	***	2.33	***
th=pronominal	2.19	***				

The forward selection (2) and backward elimination (3) approaches lead to almost the same regression model. The only difference is that in the backward model, the discourse givenness of the recipient is included as a main effect, while it is included as an interaction with medium in the forward model. Both indicate that the choice for the double object construction is more likely when the recipient has been mentioned previously in the discourse (and is thus *given*). In the forward model, this effect is a little stronger in writing than in speech.

The animacy of the recipient is only found significant in the model obtained by keeping the significant variables (1). The other differences between the two stepwise models and this model are likely to be caused by the fact that the information contained in the variables shows considerable overlap. Pronominal and definite objects are also often discourse given. A significant effect for the one variable may therefore decrease the possibility of regarding the other as significant. This is exactly what we see: the model obtained through the two

nificant decrease in accuracy in 10-fold cross-validation for the mixed models. After publication, we discovered that the `R` code we employed ignored the random effect values when applying a mixed model to test data. In this chapter, we rectified this problem, which changed the prediction accuracies (but not the models themselves).

stepwise approaches contains a variable denoting the givenness of the theme but none describing its pronominality or definiteness, while it is the other way around for the model with the significant variables from the full model.

The model obtained by keeping the significant variables in the full model also contains one interaction, namely that between medium and a pronominal theme. The main effect (without medium) is also included, but it shows the opposite effect. When the theme is pronominal, speakers tend to use the prepositional dative construction (coefficient 2.15). This effect seems much less strong in writing (remaining coefficient 2.15 - 2.01 = 0.14).

What remains unclear, is which of the three models is more suitable for explaining the British English dative alternation. Seeing the differences between the significant effects in the three models we found, it seems that the models are modelling the specific data set rather than the phenomenon. A probable cause is that the mixed models are too complex to model a data set consisting of 930 instances. In the next section, we apply the three approaches to build simpler models, namely without the random effect.

2.5.2 Models without a random effect

The model fit and prediction accuracy for the models without a random effect can be found in Table 2.6.

Table 2.6: Number of variables selected, baseline accuracy, concordance C, and accuracies ± their confidence intervals (for model fit) or two times the standard deviations (for 10-fold cross-validation) for the regression models without a random effect

			model fit (train=test)		*10-fold cv*
selection	#variables	baseline	*C*	accuracy	av accuracy
1. significant	6	0.723	0.938	0.878 (±0.021)	0.872 (±0.083)
2. forward	7	0.723	0.943	0.878 (±0.021)	0.876 (±0.089)
3. backward	8	0.723	0.946	0.882 (±0.021)	0.876 (±0.057)

The estimates of model fit concordance *C* and accuracy are considerably lower than the values obtained with the mixed models (Table 2.4). On the other hand, the models without a random effect generalise well to new data: the prediction accuracy in 10-fold cross-validation is very similar to the model fit accuracy (when training and testing on all instances). The prediction accuracies reached in 10-fold cross-validation are significantly better than those reached with the best mixed model ($p < 0.001$ for the three regular models compared to the backward mixed model, following the Wilcoxon paired signed rank test).

Table 2.7 shows the significant effects in the models without random effect. Again, the directions of the coefficients are the same across the three models, but they disagree on the significance of the variables. Three variables are selected in all three approaches: the person of the recipient (local or non-local), the pronominality of the recipient, and the concreteness of the theme. The latter two were not selected at all in the mixed-effect approach of the previous section. Three more variables have significant effects in two of the three models. According to all three models, speakers tend to use the double object construction when the theme is longer than the recipient. The forward selection model (2), however, shows that the effect of length difference is especially strong in speech. As for the mixed model in the previous section, the forward selection has selected the interaction between the medium and the discourse givenness of the recipient. Writers are thus more likely to choose the double object construction when the recipient has recently been mentioned in the text, than when the recipient is newly (re)introduced.

Table 2.7: Coefficients of significant effects in regression models (without random effect), trained on all 930 instances, *** p<0.001 ** p<0.01 * p<0.05. The (negative) effects above the horizontal line draw towards the double object construction, and the (positive) effects below it toward the prepositional dative construction.

Effect	1. significant		2. forward		3. backward	
length difference	-1.73	***			-2.00	***
length difference, medium=spoken			-2.35	***		
length difference, medium=written			-1.71	***		
rec=definite			-1.01	**	-1.15	***
rec=given, medium=written			-0.66	*		
rec=local	-1.22	***	-0.94	**	-1.15	**
rec=pronominal	-1.35	***	-0.88	**	-1.25	***
verb=abstract, medium=written					-0.99	*
verb=transfer, medium=spoken					-1.04	*
verb=transfer, medium=written					-1.32	*
(intercept)			0.82	**	1.56	**
th=concrete	1.33	***	1.48	***	1.63	***
th=definite			1.58	***	1.16	***
th=given	1.48	***			0.98	**

The semantic verb class is only selected in the backward elimination. In the literature (cf. Pinker, 1989), it is argued that the prepositional dative construction is especially used to express a change of place (moving the theme), and the double object construction a change of state (possessing the theme). In this

perspective, we would expect instances with a *transfer of possession* to be in the prepositional dative construction (*give a book to you*), and instances with *abstract* meanings in the double object construction (*give you moral support*). This is also what Bresnan et al. (2007) found for spoken American English. In the backward model, however, the effect is the opposite: a transfer of possession is more strongly drawn towards the double object construction than an abstract meaning. The problem here is that these two semantic verb classes depend largely on the concreteness of the theme (Pearson correlation = 0.739), a feature that has been selected in all three models in Table 2.7. When the semantic verb class is *transfer of possession*, the theme is very likely to be *concrete*. The backward model thus seems to compensate the positive coefficient of concreteness (1.63) by given a negative coefficient to the semantic verb class (e.g. -1.32 for *transfer of possession* in writing). The resulting effect is still directed at the double object construction (remaining coefficient 1.63 - 1.32 = 0.31), but it is not very strong. In Section 2.3, we saw that only *pay* and *send* showed different biases towards one of the two constructions in different verb senses. It seems that the biases are mostly due to the verb (see also Gries & Stefanowitsch, 2004) and the concreteness of the theme, and not so much to their semantic verb classes *abstract*, *communication* and *transfer of possession*.

2.6 Discussion and conclusion

In this chapter, we built regular and mixed (i.e. containing a random effect) logistic regression models in order to explain the British English dative alternation. We used a data set of 930 instances taken from the ICE-GB Corpus, and took the explanatory factors suggested by Bresnan et al. (2007), as introduced in Chapter 1. The regular and the mixed models were constructed following three different approaches: (1) providing the algorithms with all 29 variables and keeping the significant ones, (2) starting with an empty model and forwardly successively adding the most predictive variables, and (3) starting with a model with all 29 features and backwardly successively removing the least predictive variables. In total, we thus have built six logistic regression models for the same data set.

Both the models with and without a random effect generalise well to previously unseen data, but the models fit the data better when verb sense is included as a random effect. The six models show some overlap in the variables that are regarded significant. These variables show the same effects as found for American English (Bresnan et al., 2007): pronominal, relatively short, local (first or second person), discourse given, definite and concrete objects

typically precede objects with the opposite characteristics. With respect to *medium*, there seem to be some differences between the dative alternation in speech and writing. Four variables were selected as interactions with medium. Only one of them, the givenness of the recipient, has been selected in more than one model (i.e. in the two forward selections).

The fact that the six approaches led to six different models could be due to the relatively small size of our data set. Later in this thesis, we therefore aim at extending our data set, employing the British National Corpus (BNC Consortium, 2007). Since manually extending the data set in a way similar to that taken to reach the current data set of 930 instances is too labour-intensive, Chapter 4 will introduce an approach to automatically extend the data set (in an approach similar to that taken in Lapata, 1999), and automatically annotate it for the explanatory features in this chapter. With the larger set, we hope to be able to model the underlying processes of the dative alternation, rather than modelling the instances that made it into the current data set (see Chapter 6).

One of the drawbacks of variable selection is that different selection methods can lead to different models (Izenman, 2008), especially in small or medium-sized data sets. Accordingly, the six methods we applied have led to six different selections of variables and thus to six different models. How can we decide which is the optimal model for our purpose? Of course, the way to approach this issue depends on the goal of a specific research enterprise. For a researcher building a machine translation system, the best approach is probably to choose the highest prediction accuracy on previously unseen data. For linguists, however, the best approach may be different. In this thesis we want to combine the explanatory features suggested in previous research and test the combination on real data. We thus have hypotheses about what are explanatory features and what kind of effect they show in isolation, but it is unclear how specific features behave in combination with others. Also, we want a model that is interpretable in the framework of some linguistic theory and that, ideally, reflects the processes in human brains. It is uncertain how (and if) we can evaluate a model in this sense. Still, despite these difficulties, using techniques such as logistic regression is useful for gaining insight in the relative contribution that different features have on the choices people make when there is syntactic variability. But contrary to what seems to be common in linguistics, researchers should be careful in choosing a single approach and drawing conclusions from one model only. Firm conclusions about mental processes can only be drawn if similar models are obtained with a number of different data sets. In addition, models derived from corpus data should be tested in other types of data, such as experimental or judgement data (as will be done in Chapters 3 and 5).

3

Feature definition: concreteness

Edited from: Theijssen, D., van Halteren, H., Boves, L., & Oostdijk, N. (2011b). On the difficulty of making concreteness concrete. *Computational Linguistics in the Netherlands Journal*, *1*, 61–77. ISSN: 2211-4009.

Abstract

The use of labels of semantic properties like 'concreteness' is quite common in studies in syntax, but their exact meaning is often unclear. In this chapter, we compare different definitions of concreteness, and use them in different implementations to annotate nouns in two data sets: (1) all nouns with word sense annotations in the SemCor corpus, and (2) nouns in a particular lexico-syntactic context, viz. the theme (e.g. *the poisonous apple* in Chapter 1) in prepositional dative and double object constructions. The results show that the definition and implementation used in different approaches differ greatly, and can considerably affect the conclusions drawn in syntactic research. A follow-up crowdsourcing experiment showed that there are instances that are clearly concrete or abstract, but also many instances for which humans disagree. Therefore, results concerning concreteness in syntactic research can only be interpreted when taking into account the annotation scheme used and the type of data that is being analysed.

3.1 Introduction

Many syntacticians commonly use labels referring to semantic properties in their research, such as animacy, imaginability, concreteness, etc. These terms have become so familiar that explicit definitions of the semantic properties implied are hardly ever provided. But when definitions are provided, these may differ between different researchers. To complicate things even further, the same definition can be instantiated in different implementations or annotation guidelines. The eventual annotations can vary with respect to the range of the values that can be assigned, and with respect to their measurement scale: e.g. binary, nominal (multiple labels that cannot be ordered on a scale from 'low' to 'high'), ordinal (multiple labels that are ordered), or interval-level scores. Finally, there is the issue of replicability. For manual annotations, it is often difficult to reach high inter-annotator agreement (e.g. Theijssen et al., 2009), casting doubt on the quality of the data. For (semi-)automatic implementations, in which one employs automatic algorithms or tools, the replicability is guaranteed, but the validity may be questionable.

Existing research directed at comparing annotation approaches often suggests ways to standardise the different labels used in different language resources (e.g. Ide & Romary, 2008), or presents methods to assess the agreement between labels assigned by different annotators using the same scheme (e.g. Artstein & Poesio, 2008) or the quality of automatically obtained labels, compared to gold standard labels (e.g. Kübler, 2007). The comparisons are made with the ultimate goal of arriving at a standard that can be used across resources and studies. In this chapter, we investigate the impact of different instantiations of the definition of 'concreteness'. Instead of trying to define some standard definition of concreteness, we want to establish how the differences between the definitions and the implementations affect the actual labels obtained (an *intrinsic* comparison). Second, we want to investigate how the outcome of syntactic research is influenced when we employ the labels obtained through the different approaches (an *extrinsic* comparison).

For tackling the first goal, we employ a data set consisting of 68,484 nouns annotated with a WordNet word sense in the SemCor corpus (Miller et al., 1993): the SemCor data set.[1] Using four (semi-)automatic approaches, we assign values for concreteness to these nouns and compare these quantitatively and qualitatively.

For the second goal, we take as a case study the English dative alternation. The logistic regression models in Chapter 2[2] show that if the direct object or

[1] We make a typographic distinction between the name of the corpus (SemCor) and the set of sense-annotated nouns (SemCor).

[2] Chapter 2 presents six different regression models for the same data set. Only the models

'theme' is **concrete** (e.g. *the poisonous apple*) speakers tend to place it *before* the recipient *him*, while it is the other way around if the theme is **abstract** (e.g. *my love*). We assign concreteness values to the themes in 619 instances (the DATIVE data set) using six different labelling approaches. The labels from the various approaches are then included in logistic regression models that predict which of the two constructions is used. We compare the models to see how the selection of an approach to annotate concreteness affects the eventual conclusions.

As will become evident in the analyses, the actual labels in the SEMCOR data and the conclusions in the syntactic study based on the DATIVE data are indeed different for the various approaches. This makes us question to what degree humans agree on the interpretation of concreteness. To address this issue, we perform a follow-up experiment in which we ask humans to rate the concreteness of noun tokens in context, without providing them with a definition.

The chapter is structured as follows: Section 3.2 presents an introduction to concreteness and a description of the (semi-)automatic labelling approaches we use. Section 3.3 addresses the intrinsic comparison on the SEMCOR data, Section 3.4 the extrinsic comparison with the DATIVE data. The follow-up experiment is presented in Section 3.5. A summary and conclusion can be found in Section 3.6.

3.2 Annotation approaches for concreteness

The distinction between *concrete* and *abstract* has been addressed in a broad range of research topics, e.g. semantics, anaphora resolution, probabilistic syntax, metaphors, word sense disambiguation, syntactic/semantic acquisition and image retrieval. Quite a lot of the literature is written from the viewpoint of the antonym of concreteness: *'abstractness'*. Spreen and Schulz (1966) explain that there are at least two definitions of **abstract**: (1) general, generic, not specific, and (2) lacking sense experience. We can thus interpret concreteness as either 'specificity' or 'sensory perceivability'. The definition of concreteness as specificity originated in cognitive and neuro-science. In linguistics, the interpretation of concreteness as sensory perceivability is most common, for example in the line of research initiated by Lyons (1977). Since there is more and more research that combines insights from cognitive science and linguistics (e.g. cognitive sociolinguistics, cf. Geeraerts, Kristiansen, & Peirsman, 2010), we include both definitions in this chapter.

According to Schmid (2000) "abstract nouns are those nouns whose deno-

without a random effect for verb include a significant effect for concreteness.

tata are not part of the concrete physical world and cannot be seen or touched. Strictly speaking, what is abstract is not the nouns themselves, but what they denote" (p. 63). What a noun denotes depends on its context. Concreteness should thus not be established for different word *types* (i.e., orthographic forms), but for the individual word *tokens* in context, since words can have several senses. For instance, the noun *table* may refer to the concrete object standing in a furniture showroom, or to the tables in this chapter. Furthermore, words in the same or similar sense can be used figuratively: when a waitress shows you a table, she may be literally showing you a concrete object (a specific *table*), but what she means is not just this table, but a place to have dinner. As a result, this *table* could be considered less concrete than the one standing in the furniture showroom.

Most theories about concreteness agree that context is essential, but not all actual labelling approaches take context into account. Projects with human annotators have usually employed a two- or seven-point scale to establish the concreteness of nouns. In the binary distinction between concrete and abstract, concrete nouns are usually defined as referring to tangible, prototypically concrete, or real existing entities. In (semi-)automatic approaches to establish the concreteness of noun senses, the use of the lexical database WordNet (Fellbaum, 1998) is quite common. One can also look up the concreteness of noun types in databases, such as the MRC Psycholinguistic Database (Coltheart, 1981).

Other researchers have developed automatic approaches that use the context to assign noun tokens to noun classes (e.g. classes such as 'building' and 'event', Thelen & Riloff, 2002). One can start out with a seed set of unequivocally concrete and abstract noun types, and use a bootstrapping process: Seek patterns in the context of concrete and abstract nouns, and use them to find new examples of concrete and abstract nouns for the seed set. The patterns found after the final iteration are used to establish the concreteness of the data that needs annotation.

We compare four different approaches to annotate concreteness. The resulting labels differ in their underlying definition (specificity or sensory perceivability), the 'noun level' serving as their basis (token, sense or type), the measurement scale (interval, ordinal or nominal) and the manner in which the labels are obtained (semi-automatically or automatically).

3.2.1 MRC: The MRC Psycholinguistic Database

In this approach, we automatically look up the sensory perceivability of a noun *type*, being an interval value between 100 and 700.

The MRC psycholinguistic database (Coltheart, 1981) contains 4,004 (pro)noun

types that are marked for concreteness with a value between 100 and 700, the higher the value, the more concrete. The scores are based on ratings of words in isolation, assigned by undergraduate students. Examples of very concrete nouns are *milk*, *tomato* and *grasshopper*, very abstract nouns are *unreality*, *infinity* and *while*. The annotation instructions were taken from Spreen and Schulz (1966): "Nouns may refer to persons, places and things that can be seen, heard, felt, smelled or tasted or to more abstract concepts that cannot be experienced by our senses" (p. 460).

3.2.2 Boots: Bootstrapping the BNC

In this approach, we automatically establish the sensory perceivability with the help of a bootstrapping approach that assigns interval values.[3] In Section 3.3.2, we will see that the approach is mostly *type*-based, despite the fact that it aims at classifying individual noun *tokens*.

We use the 100-million-word British National Corpus (BNC Consortium, 2007) and parse it with the FDG dependency parser developed at Connexor Oy (Tapanainen & Järvinen, 1997). This parser has a word class (part-of-speech) accuracy of 99.3%, and it reaches a precision of 93.5% and a recall of 90.3% for linking subjects and objects.[4] From the parses, we extract all words that the parser marked as nominal head (*%NH*) and noun (*N*). We only keep those instances that have also been tagged as a noun in the BNC (*NN**), excluding for example all proper nouns. For each remaining instance, we find the lexico-syntactic patterns in which the heads appear and save them as features. The features are thus the direct dependency relations that exist between the noun phrase head and other words in the dependency parse. So, for the sentence *The major impact is yet to come*, the features for the head *impact* are *m:subj;be* (it is the subject of its mother *be*), *d:det;the* and *d:attr;major* (its daughters are the determiner *the* and the attribute *major*). Notably, the head noun itself is not included in the features. The resulting *BNC set* contains 17,708,616 tokens (170,893 different noun types).

For the initial seed set, we take the examples in Garretson (2003), consisting of 27 prototypically concrete nouns (e.g. *apple*, *door* and *knee*) and 10 prototypically abstract nouns (e.g. *air*, *current* and *molecule*). We find all instances of these nouns in the BNC set and label them **concrete** or **abstract** accordingly. Next, we train a pattern discovery algorithm on this 'labeled set' inspired by the procedure in Thelen and Riloff (2002). We aim to identity fea-

[3] In theory, the values range from $-\infty$ to ∞, but in practice, the range varies per data set: -0.50 to 0.85 for SemCor and -0.23 to 0.18 for Dative.

[4] These figures were established on texts from the Maastricht treaty by Connexor Oy in December 2005, after which only minor changes have been applied to the parser.

tures F which can act as 'concrete markers', i.e. the F with sufficient precision in suggesting concreteness and high enough frequency to be of use. We first count the number of concrete instances in the labeled set that have feature F and divide this number by the total number of concrete instances in the labeled set: $PropF_C$. The same is done for the abstract instances, yielding proportion $PropF_A$. We then check whether $PropF_C \geq w \cdot PropF_A$. The w is a strictness weight that makes the procedure stricter or more lenient, which we set to 9. Also, we demand that F occurs at least 50 times.[5]
Next, the strength of a concrete marker is established with:

$$Str_C = (\frac{PropF_C}{PropF_C + PropF_A})^2 \ . \tag{3.1}$$

The division of the proportions on the right-hand side of the equation is squared in order to enhance the relative value: strong markers are made relatively even stronger and weak markers relatively even weaker. After calculating the strengths of the concrete markers, we use the same approach to find features that are 'abstract markers', assigning them strength Str_A.

The instances in the unlabeled set are assigned a score for concreteness by adding all strengths Str_C for the concrete markers present, and subtracting all strengths Str_A for the abstract markers present. We next group all instances of the same noun and take the average of the scores assigned to them in order to find new seed nouns. The nouns are ranked according to this average score, and we add the top 100 nouns (with an average score above 0.001) and the bottom 100 nouns (with an average score below -0.001) to the list of concrete and abstract seed nouns, respectively. With the new seed set, we start over again, looking for new patterns and new seed nouns. After 100 iterations, we stop and use the final set of patterns and the final set of seed nouns, both including the concreteness scores they yield.

Using the FDG parser and the set of markers and seed nouns, we can assign a score to new instances. If no score can be assigned to the noun *token* because no concreteness or abstractness markers are present, we check whether the noun *type* (i.e. the lemma) is present in the final set of seed nouns. If so, we assign the average score corresponding to that seed noun.

3.2.3 WN-Hier: The hierarchy level in WordNet

In this approach, we semi-automatically establish the specificity of a noun *sense*, assigning an ordinal value between 0 and 16. Different tokens with the same sense always receive the same score, which is different from a true

[5] Both numbers were established by trial and error, manually monitoring the selection of new seeds.

token-based approach that takes into account the context of the individual token.

We follow Changizi (2008), who established the *specificity* of a noun with a particular sense by counting the number of hypernyms above it, using the WordNet hierarchy (Fellbaum, 1998). For instance, *bullock* is very concrete, having the maximum hypernym level of 16, while *entity* is very abstract, with the minimum hypernym level of 0. This approach is semi-automatic: The word sense is manually assigned on the basis of the context, and we use this sense to establish its hypernym level automatically by looking it up in the manually designed WordNet. It is the only approach in this chapter that uses the definition of specificity, not sensory perceivability. We thus expect this approach to differ the most from the others.

3.2.4 WN-Phys: Physical entities in WordNet

In this approach, we semi-automatically establish the sensory perceivability of a noun *sense*, assigning one of the two nominal values 0 and 1.

Again, this is not at the type or token level, but at sense level. We follow the approach in Xing, Zhang, and Han (2010): We automatically check whether the noun sense is traced back to *physical entity* in WordNet. If so, it is labelled 1, and otherwise 0. Again, this approach is semi-automatic since the word senses are found manually, and employ the manually established WordNet.

3.3 Intrinsic comparison: SemCor

3.3.1 Data

The SemCor corpus contains manually assigned WordNet word senses for all 88,058 noun phrases in 186 texts taken from the Brown corpus. Since we also want to apply the Boots approach, we parse the sentences with the FDG dependency parser and keep only those instances that the parser marked as being a nominal head. The result is the data set SemCor consisting of 68,484 instances.

3.3.2 Method and results

We apply the four approaches for concreteness labelling to the 68,484 instances in SemCor. For 44,395 instances, the noun is present in the MRC Database, which means there is a missing value for Mrc in 24,089 instances. These instances are not included in the evaluations with Mrc. There are also missing values for Boots: For 8,835 instances, no score could be assigned because there was no concrete or abstract marker present, and the noun itself was

not present in the seed list. All evaluations of Boots are thus based on the 59,649 instances for which we could assign a concreteness score. Only 5,099 of these scores are based on the presence of abstract or concrete markers in the individual *tokens*, the rest is assigned a score by looking up the concreteness score of the noun *type* in the list of seed nouns. Apparently, the set of markers is too sparse to enable the classification of the noun tokens in their lexico-syntactic context.

The concreteness scores found are compared with the help of the Spearman rank correlation coefficient, cf. Table 3.1. For comparisons with missing values (i.e. all comparisons except that between WN-Hier and WN-Phys), we only use the instances without missing values (42,120 for the comparison between Mrc and Boots). The highest correlation (0.64) is between Mrc and Boots, which could be the result of the fact that both are mostly *type*-based. The correlation between Mrc and WN-Phys is only slightly lower (0.60). WN-Hier differs most from the other three approaches, with all correlations below 0.30.

Table 3.1: Spearman correlations between the different labelling approaches for the SemCor data set. The corresponding p-values are all < 0.001.

	Mrc	Boots	WN-Hier	WN-Phys
Mrc	1.00	0.64	0.29	0.60
Boots		1.00	0.12	0.47
WN-Hier			1.00	0.17
WN-Phys				1.00

To better understand the scores, we compared the average values of the senses in WordNet's 26 noun classes. Nouns in the classes 'animal' (e.g. *hen*), 'food' (e.g. *milk*), 'artifact' (e.g. *door*), 'body' (e.g. *arms*) and 'substance' (e.g. *water*) mostly received high concreteness scores in all four approaches, and those in the classes 'relation' (e.g. *relationship*), 'cognition' (e.g. *will*) and 'attribute' (e.g. *consequence*) mostly low scores. The approaches assigned medium or varying scores to nouns in the classes 'act' (e.g. *war*), 'group' (e.g. *people*) and 'phenomenon' (e.g. *daylight*).

Boots is exceptional in assigning nouns with noun class 'time' (e.g. *February*, *minute*) and 'quantity' (e.g. *ton*, *inch*) very high concreteness scores, while they are considered (relatively) abstract by the other three approaches. Apparently, somewhere in the bootstrapping process, nouns denoting time or quantity have been included as seed nouns. As a result, time and quantity specific contexts were selected as markers of concreteness, leading to the selection of even more time and quantity nouns as seeds.

For WN-Hier, nouns that have the noun class 'object' (e.g. *soil*, *unit*) receive relatively low scores because they are located relatively high up in the WordNet hierarchy, but relatively high concreteness scores in the other three approaches. For nouns with the noun class 'feeling' (e.g *trouble*) it is the other way around: WN-Hier considers them rather concrete because they are deeper in the WordNet hierarchy, while they are considered abstract in the other three approaches.

In order to discover how individual noun types are treated in the four approaches, we determined the 100 words with the highest scores for concreteness and abstractness.[6] For the words with multiple senses we averaged concreteness/abstractness over the senses.[7]

All four approaches agree that the noun *knowledge* is abstract. In addition, three of the four approaches have placed the following nouns in the bottom 100: *ability*, *approval*, *attitude*, *confidence*, *distinction*, *freedom*, *hatred*, *importance*, *indication*, *individualism*, *morality*, *motive*, *past*, *philosophy*, *quality*, *relationship*, *responsibility*, *security*, *sentiment*, *theory*, *understanding* and *weakness*. Eight nouns have been placed in the top 100 (very concrete) by all four approaches: *cigarette*, *coat*, *grass*, *hat*, *jacket*, *sheep*, *shirt* and *tree*.

3.3.3 Discussion

Our comparison showed that the concreteness labels assigned by the four approaches vary considerably. We found no clear effect of the noun level (sense, type), the measurement scale (binary, ordinal or interval), or the annotation manner (semi-automatic or automatic). The differences we found were mostly caused by the definition used: As we would expect, 'specificity' and 'sensory perceivability' are quite different concepts. The diverging definition of WN-Hier ('specificity') made it differ greatly from the other approaches: It showed the lowest correlation with the other approaches, and the largest differences in the treatment of individual noun types and noun classes. While the other three approaches on average considered nouns denoting objects rather concrete and nouns denoting feelings rather abstract, this was the other way around for WN-Hier.

An analysis of Boots showed there is a risk in starting with a list of typically concrete and abstract nouns and using a bootstrapping approach to discover new concrete and abstract nouns on the basis of the lexico-syntactic context: Nouns denoting time and quantity were considered very concrete by Boots.[8]

[6] Ranking seems problematic for the binary approach WN-Phys, but since labels are sense-based they can have different values for the same noun type. The average score is thus not necessarily 0 or 1.

[7] If number 100 is part of a tie, we include all nouns with the same value.

[8] Similar effects occurred when we experimented with different initial seed sets.

Moreover, the concrete and abstract markers learned from the BNC were hardly present in the SemCor data, which means that the fall-back option, the score of the nouns in the seed list, determined the concreteness score in most cases. As a result, most instances were not assigned a concreteness score according to their lexico-syntactic context (the individual *token*), but according to their lemma (the noun *type*).

The automatic look-up Mrc has the most in common with the other three approaches. This is rather surprising, since it is the only approach that does not take into account the context at all. Apparently, the concreteness scores in the MRC database were based on the word sense that is the most frequent, minimising the effect of ignoring the intended sense. Also, different senses of the same noun can be similar with respect to concreteness. For instance, *arms* in the sense of weapons is arguably equally concrete as *arms* in the sense of the body parts. It thus seems that the lack of sense disambiguation in the MRC Database has little effect on the actual concreteness labels. The same is probably true for Boots, which is also mostly based on the noun types, and most similar to Mrc according to the correlation coefficients. However, because the MRC Database was designed for a different purpose (providing test items for psycholinguistic experiments), its coverage is problematic: over 24,089 instances could not be annotated with Mrc. The number of unclassified items is much smaller for Boots: 8,835.

3.4 Extrinsic comparison: Dative

In this section, we investigate how the choice for a concreteness labelling approach affects our conclusions in a syntactic study: the English dative alternation.

3.4.1 Data

We use the data set of Chapter 2, containing 930 instances of the dative alternation that were extracted from the one-million-word syntactically annotated ICE-GB corpus (Greenbaum, 1996). The ICE-GB corpus contains written and spoken British English in various genres. The data set contains manual annotations for the explanatory features introduced in Chapter 1 (see Appendix for the annotation manual). From the data set, we select those instances that were automatically parsed as an instance of the dative alternation by the FDG parser (Tapanainen & Järvinen, 1997), and were manually approved by the first author (see Chapter 4). The resulting data set (Dative) consists of 619 instances: 499 (80.6%) double object and 120 (19.4%) prepositional dative constructions.

For annotating the concreteness of the theme,[9] we employ the four approaches in the previous section. Moreover, we include two additional approaches to establish the concreteness: the manual approach used for the annotation of the original data set (MANUAL) and an adapted bootstrapping approach (BOOTS-OBJ). Both approaches are described below.

MANUAL: **Prototypically concrete**

The sensory perceivability of a noun token is manually established, assigning one of the two nominal values 0 and 1. It is the concreteness feature as it is already annotated in the data set of Chapter 2, on the basis of the annotation manual in the Appendix. It follows Garretson (2003), who deems a noun concrete if it refers to a prototypically concrete object: "The rule of thumb to apply is that we want to code as 'concrete' only *good* examples of concrete things" (5.6.7). All nouns that fit this description are given value 1, all others value 0. Remember that the examples in Garretson (2003) were also used as the initial seed nouns in BOOTS. As MANUAL is established by hand, the full context has now been taken into account by the human annotator, as opposed to only the direct lexico-syntactic dependencies in BOOTS.

BOOTS-OBJ: **Bootstrapping direct objects in the BNC**

In Section 3.3, we saw that BOOTS is mostly a type-based approach because the lexico-syntactic markers are too sparse to allow token-based classification. For this reason, we apply this bootstrapping method to a subset of the original set extracted from the BNC, containing the 837,755 noun tokens (31,345 noun types) that, according to the FDG parse, are the direct object of one of the 76 'dative verbs' in Chapter 4.[10] In this way, the set of patterns and seed nouns obtained are more likely to occur in DATIVE. The values assigned range from -0.34 to 0.19.

Since the DATIVE data set contains pronouns, we first manually establish the antecedents of the pronouns, if possible, and replace them by the head lemmas of their antecedents. The two WordNet-based approaches, WN-HIER and WN-PHYS, require additional manual annotation: we manually assign a WordNet sense to the theme.

As also found for the SEMCOR set in Section 3.3, a substantial proportion of the DATIVE nouns are missing in the MRC Database: The theme head is present

[9] As mentioned in Chapter 1, the concreteness of the recipient (*him*) is not researched, because it is highly imbalanced (there is a strong bias towards concrete recipients).

[10] Because of the smaller size of the bootstrap set, w is set to 5, and minimum marker frequency to 10.

in the MRC Database for 436 instances, so 183 instances have a missing value. We were able to assign a concreteness score to 546 instances with Boots and 475 instances with Boots-Obj. WN-Hier and WN-Phys also suffer from coverage issues: We were unable to find the intended word sense in WordNet for the theme head of 43 instances.

3.4.2 Method

Using the feature values, we establish a regression function that predicts the logarithm of the odds that the syntactic construction S in clause j is a prepositional dative. The prepositional dative is regarded a 'success' (with value 1), while the double object construction is considered a 'failure' (0). The regression function is:

$$\ln odds(S_j = 1) = \alpha + \sum_{k=1}^{K} (\beta_k V_{jk}) \ . \tag{3.2}$$

The α is the intercept of the function. $\beta_k V_{jk}$ are the weights β_k and values V_{jk} of the K variables k.[11] The optimal values for the function parameters α and β_k are found with the help of Maximum Likelihood Estimation.[12]

We employ nine features describing characteristics of the theme and of the recipient (e.g. *Snow White* in Chapter 1), taken from Table 1.1 in Chapter 1: the Animacy of the recipient (*Rec = anim*), the Definiteness of the recipient and theme (*Rec = defin*, *Th = defin*), the Discourse Givenness of the recipient and the theme (*Rec = given*, *Th = given*), the Pronominality of the recipient and the theme (*Rec = pron*, *Th = pron*), the Person of the recipient (*Rec = 1st/2nd*), and the Length Difference between the theme and the recipient (*Len dif th-rec*).[13] We also include a feature for the Medium, indicating whether the construction appeared in spoken or written data (*Med = wr*), and the six features for the Concreteness of the theme, obtained with the six different labelling approaches.[14] We build six separate regression models, each with one type of concreteness and the remaining ten features.

For all approaches except Manual, we have to deal with the missing data. We follow the standard procedure: All instances for which the concreteness

[11] We should note that the regression function treats the ordinal feature WN-Hier as an interval feature.

[12] We use the function `lrm()` in `R` (R Development Core Team, 2008).

[13] Length Difference is defined as the log of the number of words in the theme minus the log of the number of words in the recipient, i.e. the log of the ratio between the two lengths.

[14] We divide the Mrc score by 100 to prevent that its coefficient will become extremely small. Similarly, we multiply the values for Boots and Boots-Obj by 10 so the coefficients will not be very large.

is not known are removed before building the regression model.[15] The model for MRC is thus built on 436 instances, and those for WN-HIER and WN-PHYS on 576. For BOOTS, we use the 546 instances for which the noun *token* has a lexico-syntactic marker (only 12 instances) or the noun *type* is present in the final seed list (534 instances). In the BOOTS-OBJ version there are markers for 79 instances, and the noun type is a seed in 396 instances, leading to a total of 475 instances for which the concreteness could be established.

3.4.3 Results and discussion

The concordance C[16] is above 0.95 for all the six models, which indicates that the models fit the data well (cf. Baayen, 2008). When training and testing on all available instances, the prediction accuracy is above 0.91 for all six models, which is significantly better than the baselines reached when always selecting the double object construction (approximately 0.80, depending on the missing data). The β-coefficients and significance levels of the features in the models are presented in Table 3.2. The Table shows that employing different instantiations of concreteness results in different regression models. The differences are not only found for *Theme concreteness* itself, but also in the other features in the model.

Concreteness in the sense of sensory perceivability seems to play a role in the dative alternation, while concreteness in the sense of specificity (WN-HIER) does not (p=0.344). It thus seems that the definition most commonly used in linguistics, sensory perceivability, is indeed more informative in our corpus linguistic study.

The implementation of sensory perceivability affects the conclusions: the Concreteness of the theme is only significant at the 0.05-level for MANUAL, WN-PHYS and MRC. This means that only the implementations with manual input resulted in a significant effect in the models. This manual step consisted either of looking up the concreteness of nouns in the manually established MRC Database (MRC), of finding the noun sense by hand in the manually designed WordNet hierarchy (WN-PHYS), or of manually assigning a concreteness value to a noun token in context (MANUAL). Neither of the two bootstrapping approaches yielding interval values (BOOTS and BOOTS-OBJ) show a significant contribution to their models (p=0.327 and p=0.478, respectively). This is rather surprising given the correlation between BOOTS and MRC we found for SEMCOR.

[15] There are alternative ways for dealing with missing values in logistic regression, based on imputation or integrating out. However, the proportion of missing values is so high that the necessary estimates might not be reliable. Therefore, we take the safe way, and omit cases with missing values.

[16] We use the function `somers2()` created in `R` (R Development Core Team, 2008).

Table 3.2: β-Coefficients in regression models with different types of concreteness; *** $p<0.001$ ** $p<0.01$ * $p<0.05$ · $p<0.10$.

concreteness	Mrc	Manual	Boots	Boots-Obj	WN-Hier	WN-Phys
#instances	436	619	546	475	576	576
Mrc	0.56 **					
Manual		1.77 ***				
Boots			-0.58			
Boots-Obj				-0.19		
WN-Hier					-0.09	
WN-Phys						0.61 *
Th = defin	0.37	0.94 *	0.79 ·	0.75	0.83 ·	0.84 ·
Th = given	1.20 ·	0.85	1.02 ·	1.18 *	1.02 ·	0.83
Th = pron	1.22 ·	0.02	2.44 **	0.97	1.36 *	1.42 *
Rec = defin	-2.39 **	-1.54 *	-1.19 ·	-1.41 ·	-1.32 ·	-1.16 ·
Rec = given	-0.79	-0.49	-0.84 ·	-0.97 ·	-0.96 *	-0.73
Rec = pron	-1.30 *	-1.21 *	-0.35	-0.68	-0.29	-0.61
Rec = anim	-0.55	-0.37	-0.19	-0.18	0.26	0.09
Rec = 1st/2nd	-0.15	-0.63	-0.82	-0.68	-0.92 ·	-0.96 ·
Len dif th-rec	-2.48 ***	-2.53 ***	-2.59 ***	-2.73 ***	-2.65 ***	-2.65 ***
Med = wr	-0.30	-0.58	-0.33	-0.24	-0.43	-0.52
(Intercept)	-1.21	0.71	0.44	1.07	1.05	0.95

The coefficients for the three significant types of concreteness are positive, meaning that when the theme is (more) concrete, speakers and writers are more likely to choose the prepositional dative construction (*I gave the book to him*), and if it is (more) abstract, the double object construction (*I gave him my love*). This is the same pattern as found in Chapter 2. The only true token-based approach, Manual, yields the strongest effect in the regression model, with respect to the significance as well as the regression coefficient. Still, despite the different noun levels used – tokens for Manual, senses for WN-Phys and types for Mrc – and the different measurement scales – binary for Manual and WN-Phys, and interval for Mrc – the effects found are similar. Apparently, the definition of concreteness used and the presence of human intervention have the most influence.

When we look at the effects found for the other features in the models, we see that Length Difference in Table 3.2 is the most stable, with a highly significant coefficient of -2.7 to -2.5 in all six models. The other effects differ in significance across the different models. The features Discourse Givenness (= *new*) and Pronominality (= *pron*) are correlated,[17] which explains the variation

[17]We decided not to solve the collinearity by (for instance) combining the features with the help of dimensionality reduction algorithms such as Principle Component Analysis. Instead, we prefer to keep in the original features, being cautious when interpreting the model.

in significance across the models: MRC and MANUAL have significant effects for the Pronominality of the recipient, BOOTS-OBJ for the Discourse Givenness of the theme and BOOTS, WN-HIER and WN-PHYS for the Pronominality of the theme. WN-HIER also yields a significant effect for the Discourse Givenness of the recipient.

The missing data also seems to have an effect on the significance levels found. In the two bootstrapping and the two WordNet-based approaches, the features for Definiteness have lost significance, and in the MRC model, only the Definiteness of the recipient remains significant.

3.5 Follow-up experiment

Sections 3.3 and 3.4 have shown that the choice for a labelling approach influences the actual labels in the eventual data and consequently the conclusions we can draw in a syntactic study based on this data. Seeing these difference, we need to ask ourselves: To what degree do humans agree about the concreteness of words in context?

3.5.1 Method

To address this question, we perform a Crowdsourcing experiment on the platform Amazon Mechanical Turk.[18] We ask US-only workers to read a passage, answer a comprehension question and then indicate the concreteness of one of the nouns in the last sentence. They are not given any definition, only the following instruction:

> Each HIT consists of 4 passages of text. For each passage, you have to perform 3 actions:
>
> 1. Read it carefully.
> 2. Answer a comprehension question about the content.
> 3. Indicate how concrete a marked word is, on a scale of 1 (very abstract) to 5 (very concrete).
>
> For instance, consider the following sentence:
>
> `Consecotaleophobia, fear of chopsticks, was more of a hassle for my Japanese wife than it was for me.`
>
> The *'chopsticks'* are very concrete (5), while *'fear'* is very abstract (1).

[18] `http://www.mturk.com`

The *fear of chopsticks* in the example is deliberately selected so that the 'chopsticks' are concrete in both definitions of concreteness (sensory perceivability and specificity), and 'fear' is abstract in both definitions. In this way, the workers can work out their own definition. Each HIT (Human Intelligence Task) consists of four text passages and is awarded by $0.10 when all four multiple-choice comprehension questions are answered correctly (to prevent cheating). Each HIT is completed by 10 different workers.

The four items in a HIT are selected from our data sets: One item is labelled relatively abstract by all approaches (an 'easy abstract' item), one item is labelled relatively concrete by all approaches (an 'easy concrete' item), one has different labellings in the approaches (a 'difficult' item), and one is not covered by the MRC database and/or WordNet, or requires anaphora resolution (a 'special' item). We create 20 HITS from SemCor and 20 HITS from Dative, leading to a total of 40 HITS, and thus 160 items. An example item from SemCor ('easy abstract'):

> `Dear Julie. Thank you for your letter of 7 March. It may be difficult to give you a backstage placement during 10-12 April but you are welcome to come in on Friday 12 April to have a look around and meet our technicians. You could also stay and watch the show on Friday evening. On the Thursday, if you wanted to, you could spend a day with the Administration team who will give you a whole view of how the theatre functions.`
>
> When can Julie spend a day with the Administration team?
>
> - on Friday evening
> - on 7 March
> - on the Thursday
>
> Rate the concreteness of ***view***.
>
> - 1 *very abstract*
> - 2
> - 3
> - 4
> - 5 *very concrete*

3.5.2 Results

For each individual item, scored by 10 different workers, we calculated the average concreteness score, together with the standard deviation. We then took the mean over the items per type and data set, as presented in Table 3.3.

Table 3.3: Mean of average score (Av) per item, and mean of standard deviation (Sd) per item. Also provided: number of items per type per data set.

	SemCor			Dative		
Type	Av	Sd	#items	Av	Sd	#items
easy abstract	1.8	0.8	20	2.0	0.9	20
easy concrete	4.6	0.6	20	4.5	0.6	20
difficult	3.1	1.0	20	3.6	1.0	20
special	2.9	1.0	20	2.6	1.2	20

Table 3.3 shows that the easy abstract items obtain average scores $\leq$ 2.0 and the easy concrete items have average ratings $\geq$ 4.5. The mean standard deviation of the easy abstract items is 0.8 for SemCor and 0.9 for Dative. It is lower for the easy concrete items: 0.6 for both data sets. The difficult and special items receive mean scores that are closer to the middle (i.e. 3.0), both with mean standard deviations of 1.0 or higher.

Looking at the individual items, we see that eight items received a score of 5 by all ten workers: *bottle*, *hat*, *heels*, *mirror*, *oxen*, *room* (all 'easy concrete' items), *milk* and *oil* (both 'difficult').[19] Items that were assigned average scores of maximally 1.5, and thus were considered rather abstract, were *attitude*, *delight*, *feeling*, *feelings*, *freedom*, *integrity*, *manner*, *uncertainty* (all 'easy abstract'), *heart* and *principle* (both 'difficult'). Some individual items show relatively high standard deviations (1.4 or higher), indicating that the workers disagree about their concreteness: *it*, *Judaism*, *species*, *stick*, *that* ('special' items), *room*, *bit* ('difficult') and *arms* ('easy concrete').

The instructions given to the workers did not include any definition of concreteness. In the 'difficult' category, there are six instances for which the concreteness score assigned by the 'specificity' approach WN-Hier differed greatly from the score given by the three other approaches (using the definition of 'sensory perceivability'): *ice*, *water*, *land*, *film*, *men* and *pond*. In all six cases, the words are (relatively) concrete in the definition of 'sensory perceivability', and relatively abstract in that of 'specificity'. The workers gave these cases average concreteness scores of 4.0 or higher, which means they focussed most on the definition of 'sensory perceivability'.

[19] The workers rated the nouns in context, but we present only the nouns for the sake of readability.

3.5.3 Discussion

The results show us two main things: First, many items are easy both for the (semi-)automatic approaches and for humans. Especially the items that are relatively concrete according to the approaches (the 'easy concrete' items) are also clearly concrete for humans (shown by the low mean standard deviation). Second, the items that lead to most disagreement among the workers are mostly also problematic for the (semi-)automatic approaches (they are mostly 'difficult' and 'special' items). In case the concreteness differs in the two definitions, humans seem to prefer the definition of 'sensory perceivability'. Note that it is impossible to say whether middle-range values in the MRC Database are due to disagreement between raters or to the fact that these words denote things that are not intrinsically concrete or abstract.

Our observations indicate that there are items that are so obviously concrete or abstract, that there is (almost) no doubt about their concreteness. There are also many instances for which the concreteness is unclear, to which different persons assign different concreteness values. Still, even if different persons have different opinions about the concreteness of some noun, the *perceived* concreteness of the individual speaker can be a factor in that speaker's syntactic choices. For these cases, averaging over the speakers may lead to loss of this potentially important information. Instead, it may be more appropriate to take into account the differences between individual language users, for example by including the speaker/writer as a random effect. Another possible solution is to treat unclear cases as missing values. In this way, the unclear cases, besides perhaps decreasing the representativeness of the models a little, will not affect the models.

3.6 Summary and conclusion

We have compared different approaches to establish the concreteness of nouns. The approaches differed in the definition used, in the scale of the values that can be assigned (interval, ordinal, or nominal), the noun level they take as basis (token, sense or type) and the manner in which the values are assigned (manually, automatically, or semi-automatically). Our goal was two-fold: First, to find out in what way the actual labels of the concreteness of nouns change when using various definitions, or different implementations of the same definition. Second, to discover in what way the conclusions in a syntactic study change, when using these approaches.

With respect to the first goal, the scores assigned to 68,848 nouns in the SemCor Corpus showed considerable variation across the four labelling approaches we employed. The labellings by the only approach that used the

definition of 'specificity' instead of 'sensory perceivability', WN-Hier, differed most from those by the other approaches. The bootstrapping approach Boots was problematic because at some point in the process, abstract nouns (denoting time and quantity) were included as concrete seeds. Moreover, the lexico-syntactic markers were too sparse: For most of the cases it was necessary to use the fall-back option of looking up the concreteness of the noun *types* in the list of seed nouns, because no concrete or abstract markers were present for the individual noun *token*. The use of the MRC database in the Mrc approach was problematic because of its coverage (it was not designed as a tool for annotating words in arbitrary texts). The fact that Boots and Mrc mostly classify noun types, ignoring the word sense and the context, seemed to have no effect. We also failed to find an effect for the measurement scale used and the manner of annotation.

We approached the second goal by taking as a case study the English dative alternation. Using a data set of 619 instances extracted from the ICE-GB corpus, we built several regression models to predict the construction used, each using a different type of concreteness as a feature. The effects of the different types of concreteness varied considerably. Concreteness defined as 'specificity' did not seem to play a role in the choice. When defined as 'sensory perceivability', concreteness only seemed to play a role when the approach included manual input, either making use of the manually established MRC Database (Mrc), manually performing word sense disambiguation with the help of the manually designed WordNet hierarchy (WN-Phys), or manually assigning a value to the noun token itself (Manual). Again, we saw that the noun level and the measurement scale used have no clear effect, although the strongest effect was found for the only true token-based approach in the present research: Manual.

The results made us wonder to what degree humans agree about the concreteness of words in context. To investigate this, we employed a crowdsourcing experiment in which we asked workers to rate the concreteness of nouns presented in context. The human ratings showed that (also) for humans, there are instances that are clearly concrete or abstract, but also many instances for which humans disagree about the concreteness. In cases where the concreteness differed in the two definitions, people seem to focus most on the definition of 'sensory perceivability'.

Our conclusion is that results concerning the concreteness in syntactic research can only be interpreted when taking into account two factors: (1) the annotation scheme used and (2) the type of data that is being analysed. With respect to the annotation scheme (factor 1), we saw that the definition used and the presence of human intervention have the strongest effect. The type of data being analysed (factor 2) is relevant mostly because of the coverage

issues of the resources we employed (MRC and WordNet), and because of the differences in the concreteness ratings of individual language users.

4

Automatic data collection

Edited from: Theijssen, D., Boves, L., van Halteren, H., & Oostdijk, N. (2011). Evaluating automatic annotation: Automatically detecting and enriching instances of the dative alternation. *Language Resources and Evaluation*. DOI: 10.1007/s10579-011-9156-x.

Abstract

In this chapter, we automatically create two large and richly annotated data sets for studying the English dative alternation. With an intrinsic and an extrinsic evaluation, we address the question of whether such data sets that are obtained and enriched automatically are suitable for linguistic research, even if they contain errors. The extrinsic evaluation consists of building logistic regression models with these data sets. We conclude that the automatic approach for detecting instances of the dative alternation still needs human intervention, but that it is indeed possible to annotate the instances with features that are syntactic, semantic and discourse-related in nature. Only the automatic classification of the concreteness of nouns is problematic.

4.1 Introduction

Much effort has been – and continues to be – put into developing corpora to provide linguists with suitable data in sufficient quantities to perform their research. Still, for many types of research the availability of data remains an issue: even when numerous corpora are available, most of these are too small and/or have not been annotated with the required information. This means that linguists often have to extend the data and, while doing so, somehow have to provide the necessary annotations. This often involves costly manual labour. It might also involve acquiring copyright for the new data, since, ideally, the additional data and annotations should be made available to other researchers: only then will it be possible to verify any results of experiments based on these data. A possible approach to creating sufficiently large sets of suitable data that can also be accessed by other researchers is to make use of already existing corpora and provide what additional linguistic information is required automatically, employing computational tools. In this chapter, we address the question: Is data that is obtained and annotated automatically suitable for linguistic research, even if the data may contain a certain proportion of errors? We investigate this by focussing on a specific linguistic task considering syntactic alternation: modelling the dative alternation.

Previous research on the dative alternation has resulted in two data sets that have been created in a way many linguists create their data sets: Researchers extracted as many candidates as possible from corpora that contain manually checked syntactic parses. All candidates were manually checked and manually annotated with the features required (Chapter 2, Bresnan et al., 2007). We employ these *traditionally established* data sets in different ways. We will use the data set of Chapter 2 as a development and analysis set: to optimise the algorithms, and to evaluate the errors made by them. We will refer to this set as Ice-Trad since the instances were extracted from the British component of the ICE corpus (ICE-GB, Greenbaum, 1996). The data set established by Bresnan et al. (2007) is next used as a separate test set, for the purpose of quantitative evaluation only. It is taken from the Switchboard corpus of American telephone dialogues (Godfrey et al., 1992), and will be referred to as Swb-Trad from now on.

The goal of this chapter is to evaluate the quality of two data sets that we extract from the same corpora, but automatically: Ice-Auto and Swb-Auto. The procedure for automatically creating annotated data sets consists of two steps: finding instances of the dative alternation, and enriching them with the desired information. Both steps will be elaborately described further on in this chapter, and they are evaluated independently in *intrinsic* evaluations. In order to establish the effect of the automatic procedure on our conclusions in

linguistic research, we also need an *extrinsic* evaluation. We therefore evaluate the suitability of Ice-Auto and Swb-Auto by building new regression models on these sets and comparing the results to the models found for Ice-Trad and Swb-Trad.

The remainder of this chapter is organised as follows: Section 4.2 give a brief description of the two traditional data sets Ice-Trad and Swb-Trad. The automatic detection of instances is next described and intrinsically evaluated in Section 4.3, the automatic annotation of these instances in Section 4.4. The extrinsic evaluations are presented and discussed in Section 4.5. A discussion and our final conclusion can be found in Sections 4.6 and 4.7.

4.2 Traditional data

As development data, we employ the traditional data set of Chapter 2, Ice-Trad. It consists of instances found in the British component of the ICE Corpus (ICE-GB, Greenbaum, 1996). The ICE-GB corpus contains spoken and written British English in various genres, as can be seen in Table 4.1. The corpus can be obtained from the Survey of English Usage.[1]

The procedure for establishing Ice-Trad was as follows. First, candidate sentences were automatically extracted from the corpus, making use of its manually checked syntactic parses. Next, all candidate sentences found were manually checked. This was especially necessary for the prepositional dative instances, since the syntactic annotation of the ICE-GB corpus does not distinguish between different types of prepositional phrases at the clause level. This means that sentences like example 1, in which the prepositional phrase is a locative, are also found, and should be filtered out manually. The resulting data set contains 930 instances in spoken and written British English. The number of instances and of different verb types in each subgenre of the corpus can be found in Table 4.1. The majority construction is the double object construction, with a relative frequency of 72.3% (672/930). With respect to medium, the proportion of instances in spoken data is highest: 60.0% (558/930).

1. Fold the short edges to the centre. (ICE-GB W2D-019 144:1)

As a test set, we employ the traditional Switchboard set (Swb-Trad), a set of 2,349 instances, being a corrected version of the original set described in Bresnan et al. (2007).[2] The Switchboard corpus consists of spoken telephone dialogues in American English (Godfrey et al., 1992) and can be obtained from the Linguistic Data Consortium.[3] For details about the extraction of this data

[1] See `http://www.ucl.ac.uk/english-usage/projects/ice-gb`.
[2] We thank Prof. Joan Bresnan for sharing this data set with us.
[3] See `http://www.ldc.upenn.edu`.

Table 4.1: Number of double object constructions (DO) and prepositional dative constructions (PD), total number of constructions (tot), and verb types (vb) per subgenre in the ICE-GB Corpus. The number of corpus samples in the subgenres is given in brackets (each containing approx. 2,000 words).

Medium	Genre	Subgenre	DO	PD	tot	vb
W (200)	Non-printed (50)	Non-prof. writing (20)	11	3	14	7
		Correspondence (30)	93	32	125	29
	Printed (150)	Academic writing (40)	19	13	32	13
		Non-acad. writing (40)	35	13	48	15
		Reportage (20)	30	18	48	18
		Instructional writing (20)	25	10	35	8
		Persuasive writing (10)	8	8	16	6
		Creative writing (20)	45	9	54	18
			266	*106*	*372*	*51*
S (300)	Dialogues (180)	Private (100)	151	49	200	20
		Public (80)	116	41	157	25
	Monologues (100)	Scripted (70)	101	33	134	22
		Unscripted (30)	26	20	46	11
	Mixed (20)		12	9	21	9
			406	*152*	*558*	*43*
Total (approx. 1M words)			**672**	**258**	**930**	**65**

set, we refer to Bresnan et al. (2007). SWB-TRAD consists of 1850 instances with a double object construction (78.8%), and 499 with a prepositional dative construction. The number of different verb types is 38.

Some verbs have a clear bias towards one of the two constructions, as can be seen in Table 4.2. In the top part, we see the verbs that show a bias towards the double object construction: *tell*, *teach*, *give*, *show*, *offer* and *send*. The bottom shows that the verb *sell* prefers the prepositional dative construction. For the verbs in the middle, the alternation differs in the two data sets (*lend*, *do*, *cause*, *pay* and *bring*) or the verb only occurs in one of these data sets (*cost*, *take*). The Table thus reveals that the two data sets were established with different conditions: SWB-TRAD includes instances with *cost* and *take*, while they were not kept as instances in ICE-TRAD.[4]

Both ICE-TRAD and SWB-TRAD have been manually annotated for the features in Table 1.1 in Chapter 1. In the current chapter, we focus on the twelve features

[4]The verb *cost* was left out because two linguists (the first and fourth author) judged that no alternation is possible. The verb *take* either occurred in prepositional dative constructions that were locative, or in double object constructions that were judged to alternate with the preposition *of*.

Table 4.2: Number of double object constructions (DO) and prepositional dative constructions (PD) for the 10 most frequent verbs in ICE-TRAD and SWB-TRAD (*cost* and *take* are not included at all in ICE-TRAD). The percentages in boldface are those that are above 50%.

	ICE-TRAD				SWB-TRAD			
	DO		PD		DO		PD	
Verb	nr	perc	nr	perc	nr	perc	nr	perc
give	377	**85.7%**	63	14.3%	1078	**85.8%**	179	14.2%
offer	32	**76.2%**	10	23.8%	20	**66.7%**	10	33.3%
send	51	**68.0%**	24	32.0%	89	**64.0%**	50	36.0%
show	43	**81.1%**	10	18.9%	46	**86.8%**	7	13.2%
teach	7	**100.0%**	0	0.0%	58	**95.1%**	3	4.9%
tell	73	**98.6%**	1	1.4%	113	**96.6%**	4	3.4%
bring	7	**70.0%**	3	30.0%	19	44.2%	24	**55.8%**
cause	5	38.5%	8	**61.5%**	8	**80.0%**	2	20.0%
cost	0	0.0%	0	0.0%	137	**100.0%**	0	0.0%
do	10	50.0%	10	50.0%	25	**92.6%**	2	7.4%
lend	9	**52.9%**	8	47.1%	2	**66.7%**	1	33.3%
pay	8	32.0%	17	**68.0%**	83	**58.9%**	58	41.1%
take	0	0.0%	0	0.0%	2	3.4%	56	**96.6%**
sell	1	8.3%	11	**91.7%**	30	40.0%	45	**60.0%**

that describe characteristics of the theme (*the poisonous apple* in Chapter 1) and the recipient (*Snow White*), presented in Table 4.3.

The manual annotations of ICE-TRAD were done by the first author, following the annotation instructions provided in the Appendix. The definitions are as close as possible to the descriptions used for SWB-TRAD (Bresnan et al., 2007). In order to establish the quality of the data sets, we had an extra human annotator annotate subsets of the data sets. For ICE-TRAD, the third author annotated 10 items that were randomly selected, after which he was provided with feedback about his annotations. After this short training session, he annotated 40 additional instances, on which κ scores were established. Only the inter-annotator agreement for Animacy of Recipient was below 0.75 (0.63). This unexpectedly low κ score was the result of only three disagreements, all concerning groups of people that could be interpreted either as institutions (being inanimate) or as groups of individuals (being animate). They have such a great impact on the κ score because there is a great bias towards animate recipients in the 40 items. For SWB-TRAD, the first author annotated a subset of 30 items. The κ scores between these annotations and the original annotations

Table 4.3: Features and their values (th=theme, rec=recipient).

Name	Feature	Values	Description
AnRec	animacy of rec	a, in	human or animal, or not
ConTh	concreteness of th	c, a	fixed form/space, or abstract
DefRec,DefTh	definiteness of rec&th	d, in	definite pron., proper name, noun preceded by definite determiner, or not
GivRec,GivTh	discourse givenness of rec&th	g, new	mentioned ≤20 clauses before, or not (new)
LenDif	length difference	-3.4-4.2	ln(words th)−ln(words rec)
NrRec,NrTh	number of rec&th	sg, pl	plural in number, or singular
PrsRec	person of rec	l, non	local (1st/2nd) person, or not
PrnRec,PrnTh	pronominality of rec&th	p, non	headed by pronoun, or not

by Bresnan et al. (2007) were 0.78 or higher for all features, which shows a high overall agreement. The individual κ scores per feature per data set will be provided later in this chapter, in Table 4.6 (being the results of the automatic feature extraction in Section 4.4).

4.3 Automatic detection of instances in a corpus

As mentioned in Section 4.1, the first step towards automatically obtaining data sets for studying the English dative alternation (Ice-Auto and Swb-Auto) is to detect instances automatically.

4.3.1 Related work

The dative alternation, together with other *diathesis* alternations,[5] has been the topic of interest for a number of researchers in the field of automatic lexicon learning, or more specifically: 'verb classification'. Their goal has been to automatically induce possible verb frames[6] from corpora (for comprehensive overviews, see Schulte Im Walde, 2009; Korhonen, 2009). Several approaches have been rather successful (e.g. Joanis, Stevenson, & James, 2008; Schulte

[5]Diathesis alternations are alternations in which verbs systematically allow a choice between two verb frames (double object, prepositional dative) to express the same semantic roles (recipient, theme).

[6]Verb frames indicate what type of arguments a given verb can take. The definition of types depends on one's goal, and can be syntactically and/or semantically motivated.

im Walde, Hying, Scheible, & Schmid, 2008; Li & Brew, 2008; Sun & Korhonen, 2009), but many challenges are still to be met (Korhonen, 2009). Only a few researchers have attempted to tackle the detection of actual instances of diathesis alternations automatically. Their work is shortly summarised below.

Lapata (1999) used the British National Corpus (BNC Consortium, 2007) to determine the frequency with which verbs occur in prepositional dative (with *to* and *for*), and double object constructions. First, she parsed the corpus with the shallow parser Gsearch (Keller, Corley, Corley, Crocker, & Trewin, 1999) and extracted syntactic patterns that were potentially relevant. Next, she used a number of heuristic rules to divide the candidate patterns into relevant and irrelevant instances. The procedure was evaluated by comparing against manual annotations. For the double object construction (3,000 manually annotated candidates), the precision of the heuristics was approximately 89.8%, while for the prepositional dative construction with *to* (994 candidates), it was 77.3%. There is no information about recall.

McCarthy (2001) used syntactic and semantic cues to find various syntactic alternations, including the dative alternation. She parsed parts of the written part of the BNC with a probabilistic chart parser and an LR (left-to-right) parser based on string analysis. Looking at the most prototypical subcategorisation frames for each verb, she found six dative verbs that occur freely with different themes and recipients: *award*, *give*, *hand*, *lend*, *offer* and *owe*. She concluded that for the detection of instances of the dative alternation (with *to* and *for*), it is sufficient to use syntactic information only.

Lapata and Brew (2004) detected semantic preferences of verbs in the BNC and used them as priors in a Naive Bayes verb classifier. They used over 5,000 manual verb classifications to test against. Although they also evaluated the performance on the individual tokens, their task is essentially different from ours: they classify the verb class of a particular instance, while we want to detect instances of a certain verb class. The same is true for Girju, Roth, and Sammons (2005). Using the annotations available in the PropBank, they used a machine learning technique to assign verb classes to instances (tokens).

Grimm and Bresnan (2009) automatically extracted instances of the dative alternation from a POS-tagged version of the Brown family of corpora (Hinrichs, Smith, & Waibel, 2007), consisting of the written American English corpora Brown (1960s) and Frown (1990s), and the written British English corpora LOB (1960s) and F-LOB (1990s). They parsed the corpora with the Stanford dependency parser and used a Python script to extract sentences with the desired syntactic pattern and a dative verb. The sentences with complex syntactic structures (e.g. passives) were filtered out. The procedure was evaluated on a small random subset of 100 sentences with the verb *give* in the Brown Corpus. For this small set, the accuracy for automatically distinguishing

datives from non-datives was 45.0%, the recall 93.8% and the precision 46.4%. Given the low precision, they manually checked all 6,759 candidates, resulting in a final set of 3,114 instances that they used for further analysis.

4.3.2 Our method for automatic instance detection

For the automatic detection of instances, we use five steps that are performed in sequence:

- Establishing a list of dative verbs
- Extracting all sentences with these verbs from the corpus
- Parsing the sentences with the FDG parser
- Extracting candidates from the parses
- Filtering the candidates with heuristic rules

In the first step, we compile a list of dative verbs. This is not a necessary step, since we could simply include all dative constructions that the syntactic parser detects. Since we plan to use the automatic approach to case detection on very large corpora in the future, it is more efficient to first make a selection of potentially relevant sentences or utterances on the basis of a list of verbs.[7] The parsing, extracting and filtering then only needs to be applied to the retrieved sentences or utterances. Steps three to five are based on approaches in previous research. We use a syntactic parser to automatically extract potentially relevant instances like McCarthy (2001) and Grimm and Bresnan (2009). The candidates are filtered with the help of linguistic rules based on those in Lapata (1999).

In step one, we consider all verbs suggested in at least two of the following linguistic resources: the dative alternation verbs in Levin's verb classification ((1993)), the prepositional dative and double object frames in VerbNet (Kipper, Dang, & Palmer, 2000), the ditransitive verbs present in the ICE-GB corpus and the TOSCA lexicon (Oostdijk, 1996), the verbs included in Bresnan et al. (2007), a list created by Johan Bos[8] and a list in an English Grammar Guide[9]. Many of the 264 verbs found are rather rare: 86 have a frequency below 1,000 in the 100-million-word British National Corpus (BNC Consortium, 2007). The occurrences of these verbs in the BNC are often in syntactic contexts that are not dative constructions. Since the eventual goal of the automatic detection of

[7] Of course this list should not be seen as static; language changes all the time, and new dative verbs emerge.

[8] Extracted from `http://www.coli.uni-saarland.de/~bos/atp/dtvs.html` (which is no longer available): *ask*, *bring*, *buy*, *call*, *consider*, *demonstrate*, *describe*, *give*, *hand*, *leave*, *lend*, *offer*, *pass*, *promise*, *provide*, *send*, *serve*, *show*, *suggest*, *teach*, *tell*.

[9] See `http://learning.cl3.ust.hk/english-grammar-guide/Verbs/Ditransitive_Verbs.htm`.

dative instances is to prevent data sparseness in future data sets, and since the verb itself is a feature in the statistical analyses, we want to exclude such low-frequency verbs.[10] This means we remove the aforementioned 86 verbs that occur fewer than 1,000 times in the BNC (e.g. *fax*).[11] Next, we manually filter out the 102 verbs that alternate with a preposition other than *to* (e.g. *cook for*) and/or that allow only one of the two constructions (e.g. *inform*). The procedure results in the list of 76 dative verbs in Table 4.4.

Table 4.4: Final list of *dative verbs*, i.e. verbs that allow dative alternation and occur at least 1,000 times in the BNC. Verbs marked with ∗ are not recognised as allowing dative verb frames by the FDG parser.

accord	cause	flick∗	lower∗	promise	serve	take
advance	charge	fling∗	make	propose∗	ship∗	teach
allocate	concede∗	forbid	offer	quote∗	shoot∗	tell
appoint∗	deal	give	owe	read∗	show	throw
assign	deliver	grant	pass	recommend∗	sign∗	toss∗
award	deny	guarantee	pay	refuse	signal∗	trade∗
bear	dictate∗	hand	permit	repay∗	sing∗	vote
bid	do	issue	play	return∗	slide∗	wish
bounce∗	drop∗	kick∗	pose	roll∗	slip∗	write
bring	extend∗	leave	prescribe∗	sell	submit∗	yield∗
carry∗	feed	lend	present	send	supply∗	

For the second step, we extract all sentences with a dative verb which occurs in the final list. If a corpus contains POS tags, most of the times they have either been checked manually (as is the case for the ICE-GB corpus)[12] or established automatically with the help of a tagger that is trained on similar material. We use the POS tags in the corpus for a first filtering: We only extract sentences if they have a dative verb that is tagged as a verb in the corpus. This filtering is left out in the evaluation on Switchboard, where we only use the plain text in the corpus.

In step three, the sentences are fed to the Functional Dependency Grammar (FDG) parser, version 3.9, developed at Connexor (Tapanainen & Järvinen,

[10] We tested the effect of verb frequency by including it as a fixed effect in regression models for Ice-Trad and Swb-Trad. In both models, the effect of verb frequency was far from significant. We therefore believe that removing the low-frequency verbs is warranted.

[11] The threshold of 1,000 is based on our observations of the list of BNC frequencies and our intuitions about the subcategorisation frames in which these verbs may occur.

[12] Actually, the leaf nodes of the syntactic parses contain information that is similar to the result of POS tagging, and these syntactic parses have been checked manually. We will refer to this information as 'POS tags' in the remainder of this chapter.

1997). The parser outputs functional dependencies that represent the structural information within the sentence. Our motivation for choosing this parser is four-fold. First, the level of detail is sufficient for our purposes, and both dative constructions are recognized. They are explicitly marked as datives, making the extraction of candidates straightforward. For most parsers, this is not the case: they either only mark explicitly the double object construction (e.g. Stanford parser, Minipar, Link Grammar), or provide no function labels at all (e.g. Charniak). Second, the FDG parser can be used 'off-the-shelf', i.e. there is no need for training prior to applying it to data. This was an important motivation because of the small size of Ice-Trad (only 930 instances, taken from a corpus of only 1 million words), which we use for developmental purposes. The Bikel parser (Bikel, 2002), which does seem to distinguish the two dative constructions, could not be employed because it needs training. The FDG parser has been developed using approx. 100 million words in various kinds of texts – news articles, technical and legal documents, literature, discussion forum texts, transliterations, etc. – aiming for general use. Third, the parser does not need large computer capacity, and is quite fast in processing large amounts of data. Fourth, initial tests with the demo version of the parser[13] showed that it was able to deal with dative constructions with various verbs, and the parser was able to deal with complex sentences. A disadvantage of the parser is that 31 of the 76 dative verbs in Table 4.4 are not in the lexicon as being dative verbs (and cannot be added as such by users).

In the fourth step, we extract candidates from the syntactic parses. The parser generates one parse per sentence. In case a word is ambiguous, all possible functional and part-of-speech (POS) tags are provided, but it is always assigned only one relation. In dative sentences, the theme (*the poisonous apple* in the example) is labelled by the parser as an object ('obj') of the verb, while the recipient (*Snow White*), or the preceding *to* in the prepositional dative variant, is recognised as its dative complement ('dat'). We save all clauses in which one dative verb has both an object and a dative.

The fifth step consists of filtering the candidates found with heuristic rules. We distinguish between two types of filtering.

First filtering

In the first filtering, we exclude candidates that have at least one of the following features: 1) the theme or recipient is a clause, 2) the clause is in passive voice, 3) the verb is imperative, 4) the theme or recipient precedes the verb, 5) the verb is phrasal (e.g. *I'll send you out that*), 6) the clause is interrogative, 7) recipient and theme are reversed with respect to the expected order (e.g.

[13] See `http://www.connexor.eu/technology/machinese/demo/syntax`.

I give to him a letter), 8) the theme is an adjective, 9) the theme or recipient is empty, 10) the clause is a fixed expression (e.g. *I'll tell you what*), 11) there is more than one verb, theme or recipient (e.g. *I gave it to her and to him*). As mentioned in Chapter 1, most of these filters are used to prevent the influence of other types of syntactic variation than those of interest in this research (passive versus active voice, declarative versus interrogative mode, the placement of adverbials, etc.). Some are used to make sure that the features we want to apply later are applicable (e.g. it is not possible to establish the concreteness of the theme if it is a clause, not a noun phrase). We use a Perl script to apply these filters automatically, making use of the automatic parses and a manually established list of fixed expressions. This list is based on the observations made during the manual checking of the data set extracted from the ICE-GB corpus (ICE-TRAD).

Second filtering

Obviously, syntactic parsing is not the easiest task, and parsers always make mistakes. This is certainly also the case for the two dative constructions, since they are often structurally ambiguous. For the double object candidates, the difficulty lies in word sequences that are difficult to split into phrases, like *the holy water* in example 2. For prepositional dative candidates, the problem is that the prepositional phrase can be either attached to the verb or the noun (e.g. *to parliament* in example 3). These problems are even worse in automatic parsing, since even candidates that are completely unambiguous for humans, are still ambiguous for the parser since it lacks world knowledge.

2. He gave the holy water.
3. They give access to parliament.

Given the fact that parsers make errors, we have a final step in which we remove candidates that have been falsely accepted due to errors in the parses. Following Lapata (1999), we formulate a number of heuristic rules to filter out these candidates. The rules we apply are based on Lapata's work and our observations of ICE-TRAD. For some of the rules we need POS tags. For ICE-AUTO, we use the POS tags available in the corpus; for SWB-AUTO, we employ the POS tags provided in the automatic parse.

For both constructions, we remove all instances where the recipient or theme lacks the presence of a pronoun or noun. In these cases, the recipient or theme instead consists of a numeral, adjective or adverb, e.g. *a hollow* in example 4.

4. She gave [a hollow]$_{Rec}$ [laugh]$_{Th}$

For the double object constructions, there are more patterns that are likely to be the result of parse errors, or that represent structures that we do not consider instances of the dative alternation. More specifically, we filter out all candidates in which

- the last word of the recipient and the first word of the theme are proper nouns (e.g. *give John Smith*)
- the last word of the recipient is a possessive (e.g. *give Mary's money*)
- the last word of the theme is a reflexive pronoun (*give it yourself*)
- the verb is *make*, and both recipient and theme are persons in WordNet (Fellbaum, 1998) (e.g. *make him king*)
- the verb is *take*, and the theme is a time noun in WordNet (e.g. *takes me an hour*)
- the recipient and theme together are likely to be one phrase (e.g. *write the professional letters*)

For the last rule, we need to establish whether the recipient and theme together are likely to be a single object. If the recipient ends in and the theme starts with at least one noun, we take the maximum sequence of 'nouns'. This sequence not necessarily consists of real nouns only, since there may be errors in the POS tags. For instance, if we use the POS tags provided by the FDG parser (as we do for the Switchboard data), the parser could recognise a dative construction in *write the professional letters*. As a result, the word *professional* is tagged as a noun while it is in fact an adjective. We filter out such word sequences by first checking if it is present in a compound dictionary derived from WordNet (following Lapata, 1999). If it is, the candidate is rejected (e.g. *holy water* in example 5).

If it is not, we use a corpus-based approach to establish the probability that the two or three words together form a single phrase (e.g. *sea water* and *priests water* in examples 6 and 7 respectively). For this, we slightly adapt the approach in Lapata (1999), using Daudaravičius and Marcinkevičiene's *gravity* measure ((2004)), as suggested in Gries (2010), instead of the log-likelihood ratio. Gravity (G) not only takes into account the token frequencies of the separate words A and B and that of the sequence A-B, but also the number of possible word types before B and after A:[14]

$$G = \log(\frac{F_{AB} * N_b}{F_A}) + \log(\frac{F_{AB} * N_a}{F_B}), \tag{4.1}$$

in which F_{AB} is the token frequency of the combination A-B, F_A the frequency of word token A, F_B the frequency of word token B, N_a the number of possible word types before B, and N_b the number of possible word types after A.

[14] The words 'types' and 'tokens' refer to the counts of unique words and of all words, respectively.

The values are based on the British National Corpus (BNC Consortium, 2007). Using this formula, we calculate the gravity between two nouns in the sequence. If the gravity is above the suggested threshold of 5.5 (Daudaravičius & Marcinkevičiene, 2004), the noun sequence is probably a single phrase and we thus reject the candidate (e.g. *give the sea water* in example 6). Else, we keep the candidate (e.g. *give the priests water* in example 7). For three-word sequences A-B-C (e.g. *priests holy water* in example 8), we first decide how to split it into two parts by establishing the gravity between the two possible pairs (A-B and B-C). The pair with the highest gravity is next used as input to the formula, together with the remaining word (A-B and C, or A and B-C).

5. He gave the holy water.
6. He gave the sea water.
7. He gave the priests water.
8. He gave the priests holy water.

For the prepositional dative construction, we exclude all instances where the recipient is a location in WordNet (e.g. *bring him to school*). Also, we remove the instances where the prepositional phrase is likely to be the complement of the theme rather than the verb (e.g. *give access to the garden*). We again employ the aforementioned gravity measure to establish this. Since the BNC contains no syntactic annotations at the level required, we first parse the BNC with the FDG parser. We then use the PP-attachment in the parses to calculate the gravity between the verb and the recipient, and between the theme and the recipient.

4.3.3 Results

We applied the approach described in the previous section to ICE-GB and Switchboard. The number of candidates found in both corpora are presented in Table 4.5. The table also shows the precision, recall and F-score of the sets after the second filtering (Ice-Auto and Swb-Auto), when comparing them to Ice-Trad and Swb-Trad.

In Ice-Auto, 62.9% (559/889) of the instances are from the spoken part of the corpus, which is not significantly different from the 60.0% in Ice-Trad ($\chi^2 = 1.47$, df= 1, $p > 0.20$). The majority construction is the double object construction, comprising 73.1% (650/889) of the instances (which again is not significantly different from Ice-Trad: 72.3%, $\chi^2 = 0.13$, df= 1, $p > 0.70$). In Swb-Auto, the proportion of double object constructions is significantly different from Swb-Trad: 81.9% (2,206/2,694), compared to 78.8% ($\chi^2 = 7.61$, df= 1, $p < 0.01$).

Table 4.5: Results of automatic case detection for both the development/analysis data (ICE) and test data (SWB).

	ICE	SWB
Number of candidates found by parser	1,674	5,087
Number of candidates after 1st filtering	1,111	3,356
Number of candidates after 2nd filtering (AUTO)	889	2,694
Number of candidates in both AUTO and TRAD	619	1,292
Precision	69.6%	48.0%
Recall	66.6%	55.0%
F-score	68.1%	51.2%

4.3.4 Discussion

When we look at the precision, recall and F-score for both data sets, we see that the scores for ICE-AUTO are much higher than those for SWB-AUTO. In general, parsers have more difficulties with spoken material than written material, because it often contains disfluencies, corrections and unfinished clauses. We also find a trend for this within the ICE-GB data: The precision for the spoken instances in the ICE data is 67.4% (377/559), while it is 73.3% (242/330) for the instances in written English ($\chi^2 = 3.13$, df= 1, $p < 0.10$).

The approach is quite successful on the development data: with the help of the filtering rules, our approach outperforms the precision reached on the Brown corpora by Grimm and Bresnan (2009): 69.6% compared to 46.1%. The recall of our approach, however, is much lower: 66.6% compared to 93.8%. Combining precision and recall, we reach an F-score of 68.1% on the ICE-GB data, which is higher than the 61.8% obtained by Grimm and Bresnan (2009). When we compare our performance on the Switchboard data to Grimm and Bresnan (2009), we see that the precision we reach (48.0%) is comparable to the precision they reach on the Brown corpora (being 46.1%). Their F-score is much higher (61.8% compared to our 51.2%), however, because of their better recall. It is clear that the spoken data in the test set (SWB-AUTO) is problematic for our approach.

Let us now consider the errors made on the development set (ICE-AUTO). In order to gain insight into the possible improvements of the approach, we manually classified the 270 candidates in ICE-AUTO that are not present in ICE-TRAD:

- 131 (48.5%) of them are found because the FDG parser incorrectly recognised a dative construction. These sentences are not part of ICE-TRAD

because the syntactic annotations of these sentences in the ICE-GB corpus do not contain a construction that could be dative. These errors are thus in fact parse errors, which we are unable to solve (at least in the scope of this thesis).

- 125 (46.3%) are found both in the ICE-GB annotations and the automatic parses. In Ice-Trad, these were filtered out automatically (using the syntactic annotations in the corpus) or manually. The procedure is different for Ice-Auto: the automatic filtering is now based on the automatically obtained FDG parses, and the manual filtering is replaced by the filtering rules. For 125 instances, the FDG parser thus should have indicated that these constructions are irrelevant, or the filtering rules should have filtered them out. In 33 sentences, the prepositional phrase indicates a location, amount, time or degree, in 22 the verb is phrasal, 17 are fixed expressions, 15 are imperatives or interrogatives, 12 have an object that is split up or incomplete, 11 have clausal objects, and 15 are irrelevant because of other reasons. Where we manually checked these properties for Ice-Trad, we have performed no such checking for Ice-Auto.
- 14 (5.2%) actually contain relevant dative constructions. Of course, manually checked annotations are also error-prone (e.g. Nancarrow & Atwell, 2007). The fact that these 14 instances are not part of Ice-Trad exemplifies this, since they are missed due to errors in the annotations in the ICE-GB corpus. Most of these instances (11) were prepositional dative constructions in which the prepositional phrase was incorrectly attached to the theme, not to the verb.

The division shows that of the 889 candidates found automatically, 256 are not relevant instances of the dative alternation. The FDG parser thus reached a precision of 71.2% on recognising the two objects in dative constructions. This precision is much lower than the general precision the parser reaches on linking subjects and objects: 93.5% on texts from the Maastricht treaty and 96.5% on foreign news texts.[15] Of the remaining 633 candidates, not all are present in Ice-Trad, but they are all instances of the dative alternation. The effect of the 256 irrelevant cases will become evident in Section 4.5.

The 311 instances not found automatically can be subdivided as follows:

- 206 (66.2%) are due to errors in the automatic parses. Of these, 10 have a verb that is listed in our list of dative verbs (Table 4.4), but which is not marked as such in the parser lexicon (e.g. *read*, indicated with the asterisk in Table 4.4). So, the fact that 31 of the 76 dative verbs in our

[15]These figures were established by Connexor Oy in December 2005, after which only minor changes have been applied to the parser.

list are not stored as such in the parser lexicon, eventually causes only 10 instances to be missed.

- 43 (13.8%) are falsely filtered out in the second filtering. In 20 the gravity measure unjustly indicated that the recipient should be interpreted as a postmodifier of the theme. In 12 the recipient is interpreted as a location on the basis of WordNet, while it is not a location in the given context. In 11 the theme or recipient does not contain (or is taken not to contain) a (pro)noun, but only a numeral, adjective or adverb. Seeing the small number of errors per type of filtering, there is only little to gain by improving the filtering.
- 38 (12.2%) are not found because the verb is not in the list of dative verbs (Table 4.4). We have not employed this list to establish ICE-TRAD; we extracted all dative constructions regardless of the verb present, and manually checked whether the candidate was relevant.
- 24 (7.7%) are falsely filtered out in the first filtering: 10 are taken to have a clausal object, 6 are interpreted as expressions, 4 are taken to have split objects, 3 are considered passives or imperatives and 1 is taken to contain a phrasal verb. As with the second filtering, the numbers are too small to make investing in improvement worthwhile.

When we compare the recall we obtained with the FDG parser (66.6%) to the recall that the FDG parser reaches on linking subjects and objects in general, we see that the general recall of the FDG parser is much higher: 90.3% on texts from the Maastricht treaty and 95.4% on foreign news texts.[16]

Counting the number of words in the sentences in ICE-TRAD, we see that the instances that were found automatically have an average length of 21.5 words, which is significantly shorter than the average length of 25.7 words for the cases we could not find automatically (t= -4.23, df= 548, $p < 0.001$). Apparently, the parser has most difficulty identifying dative constructions in sentences that are relatively long and thus, presumably, more complex. Also, the division of the two construction types (double object and prepositional dative) differs significantly between the instances we have found in the ICE corpus and those we have missed. Of the cases we found automatically, only 19.4% is a prepositional dative, while this is true for 44.4% of the cases we have missed ($\chi^2 = 63.2$, df= 1, $p < 0.001$).[17] The attachment of prepositional phrases (PP-attachment) is a common problem in automatic parsing (e.g. Agirre, Baldwin, & Martinez, 2008). It is therefore not surprising that the parser has more

[16] Again, these figures were established by Connexor Oy in December 2005.

[17] As we will mention in Section 4.5, the difference in the distribution of the two constructions between ICE-AUTO and ICE-TRAD will not influence the extrinsic evaluation, because logistic regression is very robust against class imbalance.

difficulties with the prepositional dative variant than with the double object construction (cf. also Lapata, 1999).

4.4 Automatic annotation of instances found

Now we have described the automatic extraction of the instances, we can move on to the annotation of the instances. We include the features introduced in Section 4.2. Previous research has already shown that they play a role in the dative alternation (Chapter 2, Bresnan et al., 2007).

4.4.1 Method

The automatic extraction of the values for the twelve features is described below. Our aim is to obtain feature values that agree with the ones selected by the human annotator in Ice-Trad. We make use of the syntactic parses produced by the FDG parser. For some features, we consult WordNet (Fellbaum, 1998). Most corpora contain POS information, so we also make use of the POS tags in the ICE-GB corpus. For the Switchboard data, we use the POS tags provided by the parser.

Animacy of recipient (AnRec)

Most researchers who have been successful in animacy classification of *English* nouns employ WordNet (e.g. Orăsan & Evans, 2007; Baker & Brew, 2008). We therefore employ WordNet as well, together with other resources. More precisely, we use three lists of animate words: (1) the nouns marked as person or animal in WordNet, (2) a list of company names found on the Internet[18] and (3) a short list of additional words, e.g. personal pronouns like *I* and *him*. Company names are thus deemed animate. Our assumption is that company names functioning as a recipient in a dative construction will mostly refer to the people working at this company (e.g. *BUPA* in example 9).

9. I mean two three weeks ago John Major made a speech to BUPA in which he said he wanted the private sector to be boosted. (ICE-GB S1B-039 64:1:C)

In this chapter, we simplify the problem of animacy classification of the recipient in two ways. First of all, we limit ourselves to the lemma of the syntactic head of the recipient, as found in the syntactic parse. Second, we classify the

[18] The company names have been extracted from `http://www.buyblue.org/alphalist.php` and `http://www.businessweek.com/1999/99_28/alphalist.htm` (which seem no longer available). All names ending in Corp, Corporation, Co, Incorporated, Inc, Holding, Group were duplicated without this ending.

different types (not the different tokens), irrespective of context. This means that we always assign the same value for animacy to recipients that have as their syntactic head the same lemma. When this lemma is present in at least one of the three lists mentioned above (ignoring upper/lower case), we classify it as *animate*. All other recipients are deemed *inanimate*.

Concreteness of theme (ConTh)

In the dative alternation, the theme can either be "prototypically concrete" (Garretson, 2003), i.e. having a fixed size or form in space (*She gave him* ***a book***), or abstract (*She gave him* ***her love***). Concepts like *love* and objects like *book* are fairly straightforward, but there are many difficult cases. Sometimes words have several senses. For instance, if a furniture salesman *shows you* ***a table***, he most likely refers to the concrete object standing in his showroom, while a researcher giving a presentation refers to the representation of information he or she has put on the slide. Furthermore, words in the same (or at least in a similar) sense can be used figuratively: when a waitress in a restaurant *shows you* ***a table***, you could say she is literally showing you a concrete object (a specific *table*), but what she means is not just a table, but a place to have dinner. In this situation, the *table* is arguably not prototypically concrete anymore.

The assumption that meaning depends on context is not new. The distributional hypothesis in Harris (1954) has led to a long line of context-based approaches in lexicon learning, many of which are semi-supervised or unsupervised. There are two main lines of research: (1) clustering semantically similar words (e.g. Rooth, Riezler, Prescher, Carroll, & Beil, 1999), and (2) extending existing lexicons through bootstrapping. In both, contextual features are used to find similarities. The clustering approach is not directly useful for us, since we want a binary, pre-defined, classification. A common method in bootstrapping is to start with a very simple lexicon, comprising a set of occurrences (tokens) of word types that are prototypical examples of the semantic class of interest (in our case concrete and/or abstract words). For this *seed set*, it is assumed that the word types have this class in (almost) all contexts, so you can use all tokens of this word type. In an iterative process, the seed set is extended with new word tokens that share properties with the tokens in the seed set. Many researchers use syntactic information for this purpose, for instance for classifying nouns into the lexical categories *building*, *event*, *human*, *location*, *time* and *weapon* (Riloff & Jones, 1999; Thelen & Riloff, 2002) or for detecting film titles (Kuijjer, 2007).

In Chapter 3, we evaluated five different (semi-) automatic approaches for establishing the concreteness of nouns for the purpose of investigating the

dative alternation. One approach used the MRC Psycholinguistic Database (Coltheart, 1981) to find the concreteness value of a noun *type* (interval scale). Two approaches found the concreteness of a noun *sense* with the help of the WordNet hierarchy: by counting the number of nodes from the sense to the root (ordinal scale, following Changizi, 2008), and by checking whether the sense was part of the *physical entity* subtree (binary value, following Xing et al., 2010). The final two approaches (both resulting in an interval scale score per noun *token*) were two variants of the bootstrapping approach in Thelen and Riloff (2002): one using all syntactic contexts, and one only dative contexts. The data used for bootstrapping was taken from the British National Corpus. The conclusion was that the first three approaches were hampered by the insufficient coverage of the lexical resources used (WordNet, MRC). The bootstrapping approaches were not very successful either: abstract nouns denoting 'time' (e.g. *minute*, *February*) and 'quantity' (e.g. *ton*, *inch*) received high concreteness scores. This shows that the selection of the seed set is not trivial, and that seeds are often not as suitable as one would expect.

Seeing the problems with the use of existing resources and with applying a bootstrapping approach, we decided to make use of our development data: Ice-Trad. We take the 619 instances in Ice-Trad that were also detected automatically, and establish a number of syntactic features for them. For each instance, we find the automatically obtained FDG parses and for the head of the theme, we extract the relations to its daughter nodes and to its mother node. Also, the POS tags and lemmas are retrieved, as well as the representation (upper/lower case, etc.). The found information is transformed into machine learning features. For instance, the features for *apple* (head of *the poisonous apple*) in the example sentence are:

- lemma of the focus word: *apple*
- upper (*U*) / lower case (*L*) and presence of non-alphanumeric characters (*S*) in the focus word: *L*
- POS tags of the focus word: *N_NOM_SG*
- relation with the mother node: *obj*
- relation with the mother node + its lemma: *obj;give*
- relation with the mother node + its POS tag(s): *obj;V_PRES_SG3*
- relation with the daughter node(s): *det*, *attr*
- relation with the daughter node(s) + its/their lemma(s): *det;the*, *attr;poisonous*
- relation with the daughter node(s) + its/their POS tag(s): *det;DET*, *attr;A_ABS*

After establishing the features for the themes in the instances, we applied various machine learning algorithms to classify the 619 instances in a 10-fold cross-validation setting. We employed Weka (Hall et al., 2009) for a number

of classifiers,[19] and libSVM (Chang & Lin, 2001) for Support Vector Machines (SVMs). For all algorithms, only the features that occur at least three times in the training data were actually employed. The best results were obtained by the SVMs.

SVMs need tuning of three hyperparameters: (1) the kernel, which we limit to linear and RBF, (2) the cost c, for which we go through a grid of 2^{-12} to 2^{10} with steps of $*2$, and (3) the gamma g (only for the RBF kernel), for which we go through a grid of 2^{-10} to 2^{6} with steps of $*2$. The optimal hyperparameters are found in a 10-fold cross-validation setting, and these are then used to build an SVM on all training data. When applying the classification to the 619 instances present in ICE-TRAD, we perform leave-one-out: the tuning is done on 618 instances in 10-fold cross-validation, the optimal settings are used to build an SVM on all 618 cases, which is then used to predict the 619th instance. When predicting data that does not overlap with the 619 instances in ICE-TRAD, we use all 619 instances for tuning and training.

Definiteness of recipient and theme (DefRec, DefTh)

For establishing the definiteness of recipient and theme, we use the POS tags in the corpus or the parse. In order to establish what is the head of the object, and with which words it occurs (i.e. which words are its daughter nodes) we always use the dependency relations in the FDG parse.

When the head occurs with a definite article, we classify it as *definite*. The same applies to a head that is, or occurs with, a demonstrative, interrogative, relative or possessive pronoun. Similarly, we consider *definite* heads that are a reciprocal, reflexive or personal pronoun, or a proper noun.

Discourse givenness of recipient and theme (GivRec, GivTh)

Automatically identifying discourse-new objects has received considerable attention of researchers working in the field of anaphora resolution. This is because the first step in anaphora resolution is recognizing which elements should be resolved, i.e. which elements actually refer to an item that has previously been mentioned (and thus is discourse-given).

Vieira and Poesio (Vieira, 1998; Vieira & Poesio, 2000; Poesio, Uryupina, Vieira, Kabadjov, & Goulart, 2004) used heuristics to establish which definite nouns are discourse-new. For example, one heuristic rule says that noun phrase heads that start with a capital (e.g. *The Iraq war*), or that refer to

[19] More specifically, we used Naive Bayes, Logistic, Multilayer Perceptron, Voted Perceptron, RBFNetwork, Ibk, AdaBoost, Bagging, SimpleMI, Jrip, DecisionTable, J48 and RandomForest.

time (*the morning*) are discourse-new. Testing on 195 definite phrases, they reached a precision and recall of 77%.

The work by Vieira and Poesio has been extended in several ways. Some have added new heuristics (e.g. Bean & Riloff, 1999). Others have extended the work to other types of nouns, not only the definite ones (e.g. Ng & Cardie, 2002; Uryupina, 2003). The use of different machine learning techniques for the task of detecting discourse-new objects has also received some attention (e.g. Ng & Cardie, 2002; Kabadjov, 2007). Most researchers have employed the corpora created for one of the Message Understanding Conferences (MUC) for training and testing. The precision and recall are generally above 80%.

In the context of anaphora resolution, establishing the discourse givenness of an object affects the decision made further on in the discourse, since *new* objects will be *given* later in the discourse. Our task is considerably easier, since we only have to establish whether the recipient or theme is discourse-given or discourse-new. We therefore developed our own, much simpler, algorithm.

The approach we take is as follows. Given the fact that indefinite objects are mostly new to the discourse, we classify all indefinite objects as *discourse new*. For definite objects, we extract the head and its attributes from the FDG parse, and take the POS tags from the corpus or the parse. Definite objects of which the head is a personal pronoun, and of which the head is preceded by a demonstrative pronoun, are labelled *discourse given*. For the remaining definite objects, we check the preceding contexts, with a maximum length of 20 clauses (i.e. until the 20th preceding word that is tagged as main verb). If the head itself, or a synonym of the head is found within this preceding context, the object is considered *discourse given*. We use the synsets in WordNet to extract the synonyms. The remaining definite objects are given the value *discourse new*.

Number of recipient and theme (NrRec, NrTh)

Again, we employ the POS tags in the corpus or the parse. We use the FDG parse to identify the head of the object, and take the number provided in the POS tag. For heads of objects that have no information about number (e.g. *you*, which can be both), we assign the default value *singular*.

Person of recipient (PrsRec)

For this feature, we simply check whether the head of the recipient is *I*, *me*, *my*, *mine*, *myself*, *you*, *your*, *yours*, *yourself*, *yourselves*, *we*, *us*, *our*, *ours* or *ourselves*. If this is the case, the recipient is *local*, otherwise it is *non-local*.

Pronominality of recipient and theme (PrnRec, PrnTh)

For this feature, we again employ the POS tags in the corpus or the parse. We extract the head of the object from the FDG parse. If the head has a POS tag for (any type of) pronoun, the object is classified as *pronominal*. If not, it is *non-pronominal*.

Length difference (LenDif)

For length difference, we use a Perl function that we also used for Ice-Trad (see the Appendix). It counts the number of words in the recipient and the theme by splitting on white space, and takes the natural log of these lengths to smoothen outliers. The recipient length is then subtracted from the theme length (thus giving the log of the ratio between the lengths). The difference with Ice-Trad is that the input strings are now not the theme and recipient as found in Ice-Trad, but as found with the automatic approach (i.e. in the FDG parses).

4.4.2 Results

We intrinsically evaluate the automatic feature extraction by comparing the features values found to a gold standard. This gold standard consists of the manual annotations in Ice-Trad and Swb-Trad. For this reason, only the instances that are present in both the traditional and the automatic set can and will be included in this evaluation (619 for the ICE data, 1,292 for the Switchboard data).

For the only feature with an interval scale, length difference, we calculate the correlation coefficient between the values in the traditional set, and those in the automatic set. For the 619 instances in Ice-Trad and Ice-Auto, the correlation is 0.825. For 442 (71.4%) of the 619 instances, the feature value is exactly the same in both data sets. In 26 (4.2%) instances, the theme and recipient are equally long in one data set, but differ in length in the other. In only 8 (1.3%) instances, the polarity differs. For the remaining 143 (23.1%) instances, the same object is found to be longer than the other, but the length differences found differ.

For the 1,292 instances in Swb-Trad and Swb-Auto, the correlation is 0.635. The length difference has exactly the same value in 851 (65.9%) instances, is zero for only one of the two sets in 97 (7.5%) instances, differs in polarity in 11 (0.9%) instances, and only differs in the size of the length difference in 333 (25.8%) instances.

For the binary features, we calculated the classification accuracy and established the proportion of the majority value. The results are shown in Table

4.6. They show us two important things. First of all, most features are biased towards one of the two values. Over 90% of the recipients are definite, for instance. An exception is the person of the recipient where the division between *local* and *non-local* is much more balanced. Second, we see that the accuracies reached are all $\geq 79\%$.

Table 4.6: The accuracy (Acc) of automatic feature extraction and the proportion of the majority class (Maj) in the traditional sets for the binary features. The fourth column indicates the κ score between the traditional and the automatic annotation of the instances in both sets. Human κ scores are provided in the fourth column.

	ICE				SWB			
Feature	Acc	Maj	κ	human κ	Acc	Maj	κ	human κ
PrsRec	1.00	0.52	1.00	1.00	0.82	0.64	0.65	0.91
PrnRec	1.00	0.77	0.99	0.95	1.00	0.87	0.98	1.00
DefTh	0.99	0.65	0.97	1.00	0.97	0.70	0.94	0.93
DefRec	0.99	0.93	0.93	0.78	0.98	0.95	0.80	1.00
NrTh	0.98	0.86	0.92	0.88	0.98	0.80	0.93	1.00
PrnTh	0.98	0.88	0.89	0.84	0.97	0.84	0.91	1.00
NrRec	0.94	0.71	0.84	0.77	0.95	0.71	0.88	1.00
GivRec	0.91	0.82	0.69	0.95	0.95	0.86	0.79	0.80
AnRec	0.92	0.90	0.68	0.63	0.91	0.87	0.55	1.00
GivTh	0.87	0.83	0.59	0.80	0.90	0.82	0.68	0.78
ConTh	0.87	0.79	0.55	0.75	0.79	0.72	0.37	0.86

Seeing that all accuracies and most majorities are above 79%, we have also established the κ statistic for inter-annotator agreement that discounts the prior probability that two annotators will agree. The two annotators in this case are the human annotator of the traditional set, and the extraction algorithm for the automatic set. Table 4.6 shows these κ scores, as well as those between two human annotators (as described in Section 4.2). Most of the κ values between the automatic and the manual annotations are quite similar to the κ scores between two human annotators. This is not the case for the features that are intuitively the most difficult (givenness, animacy and concreteness); they result in lower κ scores. For the other features, the κ scores are all above 0.65.

4.4.3 Discussion

When looking at the results for the ICE data (the development/analysis data), we see that for the animacy of the recipient, the κ score between the automatic extraction and the human annotations is very similar to that between

two humans. Apparently, the simplifications in the automatic extraction have not influenced the quality of the extraction, and the resources we employed are quite reliable. One of those resources is WordNet. Of course, some noun head lemmas may have several senses and therefore occur not only as animal or person, but also in a different noun class. This is the case for 12 of the 47 incorrectly classified instances: *authority* (twice), *dealer*, *face*, *man*, *master*, *mother*, *opposition*, *party*, *plant*, *subject* and *world*. With respect to simplifications, remember we had two of them: (1) limiting ourselves to the lemma of the syntactic head of the recipient, and (2) classifying the different types (not the different tokens), irrespective of context. An analysis of the 47 misclassifications reveals that only five are caused by the first simplification: four are animate but were labeled inanimate (*those who...*, *any of...*, *the rest of...*, *Mr ...*) and one the other way around (due to a parse error in *the former Deputy Prime Minister's words*). The second simplification leads to sixteen errors in Ice-Auto. Three are nouns that are inanimate in the given context, but were labelled animate automatically: *world*, *nation* and *face*. The rest are incorrectly labelled inanimate, e.g. *few*, *it*, *jury* and *panel*.

The lowest κ scores between the automatic and manual annotations are for the concreteness of the theme, in both Ice-Auto and Swb-Auto. When we look at the 83 cases that are different in the ICE data, we see that fourteen are pronouns (*it*, *some* and *that*). It is not surprising that pronouns are difficult for the automatic approach; it depends even heavier on the context, since it needs to resolve to which antecedent the pronoun refers. In addition, there are twenty cases where the theme is something made of paper, e.g. *picture*, *piece of paper*, *card*, *voucher*. These are concrete in the manual annotation, but labelled abstract in the automatic approach. Apparently, these types of themes often share contextual properties with themes that are abstract.

Surprising is the rather low κ for the person of the recipient for the Switchboard data: 0.65. A quick look at the discrepancies shows that the annotations in Swb-Trad are more semantic in nature (whether the recipient is really physically part of the discourse), while we used a more syntactic definition (whether the recipient is in first or second person) in Ice-Trad and the automatic approach. Almost all of the differences were caused by the generic use of *you* (e.g. *it gives you energy*), which was labelled 'non-local' in Swb-Trad, and 'local' by us.

4.5 Extrinsic evaluation: Using the data in logistic regression models

The intrinsic evaluations in the previous sections have shown that the automatic detection of instances of the dative alternation may need improvement, but that the annotation of these instances seems promising. In order to establish the effect of the automatic procedure on our linguistic research, we need an extrinsic evaluation. The use of extrinsic evaluations is quite common in the field of Natural Language Processing (NLP), where the quality of the automatic annotations are tested in NLP applications like machine translation (e.g. Bod, 2007) and question answering (e.g. Theijssen, Verberne, Oostdijk, & Boves, 2007). Researchers in NLP now even question the use of *intrinsic* evaluations (e.g. Poibeau & Messiant, 2008). For our extrinsic evaluation, we build a logistic regression model on the automatic data (ICE-AUTO and SWB-AUTO), and compare the effects in the model to a model built on the traditional data (ICE-TRAD and SWB-TRAD).

4.5.1 Method

Previous research has indicated that over 90% of the dative alternation can be correctly predicted with a logistic regression model that combines the features introduced previously (Chapter 2, Bresnan et al., 2007). More information about multivariate techniques such as regression can for instance be found in Izenman (2008).

As discussed in Chapter 2, there are at least six ways to build a logistic regression model for the dative alternation. One can choose between a mixed model, i.e. a model with a random effect, and a model without such an effect. Seeing the verb biases we presented in Table 4.2, we want to include verb as a random effect.[20] The second choice we have to make is the manner of feature (or *variable*) selection. Researchers have employed at least three different approaches to feature selection: (1) first building a model on all available explanatory features and then removing those that do not show a significant contribution (e.g. Bresnan et al., 2007), (2) sequentially adding the most explanatory feature (forward), until no significant gain is obtained anymore (e.g. Grondelaers & Speelman, 2007), and (3) starting with a model containing all available features, and (backward) sequentially removing those that yield the lowest contribution (e.g. Blackwell, 2005). Comparing all three options for the two data sets is beyond the scope of the present chapter. Seeing that our research is the closest to that in Bresnan et al. (2007), we follow their approach. We will thus build only one type of model: a mixed model with

[20] We use the function `lmer()` in the `lme4` package in `R` (R Development Core Team, 2008).

verb as a random effect, building it on all features, then removing all features that are not significant, and building a new model with only those features.

The ICE data sets differ from the Switchboard data sets in the sense that they contain both spoken and written material. Following Chapter 2, we include medium (spoken or written) as an additional feature, and add all interactions of medium with the twelve features of the previous section. This leads to a total number of 25 features. When removing non-significant main effects, we never remove those that are part of a significant interaction. For the Switchboard data, containing only spoken material, we include only the twelve main features.

4.5.2 Results for the ICE data

We build two regression models for the ICE data: (1) a model built on the 930 instances in ICE-TRAD, and (2) a model built on the 889 instances in ICE-AUTO. The model quality of these models (and an additional one that will be introduced later in this section) can be found in Table 4.7. The models fit the data well: the prediction accuracy is over 87%. This is significantly better than the majority baseline of always selecting the double object construction. Also, the concordance C is above 94% for all three models. In a 10-fold cross-validation setting, the regression models show only a slight decrease in prediction accuracy and concordance *C*, which means there is hardly any overfitting.

Table 4.7: Prediction accuracy and concordance *C* for the model fit (Acc, *C*) and in 10-fold cross-validation (av Acc, av *C*), for ICE-TRAD, ICE-AUTO and ICE-SEMI. The majority baseline and the number of instances are also provided.

Data set	Majority	N	Acc	av Acc	st.dev	*C*	av *C*	st.dev
ICE-TRAD	0.723	930	0.915	0.896	0.036	0.973	0.962	0.016
ICE-AUTO	0.731	889	0.880	0.871	0.045	0.947	0.933	0.025
ICE-SEMI	0.791	633	0.930	0.918	0.055	0.969	0.954	0.035

The results in Table 4.7 give an indication of the quality of the regression models. For the qualitative evaluation, we inspect the significant effects in the two models, shown in Tables 4.8 and 4.9.

Four of the five significant features in the traditional model are also found to be significant by the automatic model (printed above the horizontal line in Table 4.9). The effect that is missing in the automatic model is the pronominality of the theme. Instead, we have a significant effect for the pronominality of the recipient. For three features that are significant in both models (givenness

Table 4.8: Significant features in the model built on the 930 instances in Ice-Trad. The coefficients β for the model fit are provided, together with the average βs in the ten separate models in the ten-fold cross-validation, and their standard deviation. Also, the p-values for the model fit are shown, as well as the average p-values in the ten separate models, with their standard deviation.

Feature	β	av β	st.dev	p	av p	st.dev.
(Intercept)	1.34	1.30	0.17	0.033	0.056	0.042
PrnTh=p	1.47	1.46	0.24	0.002	0.007	0.007
GivTh=non	-1.97	-1.98	0.15	0.000	0.000	0.000
LenDif	-2.11	-2.12	0.04	0.000	0.000	0.000
AnRec=in	0.77	0.77	0.16	0.037	0.068	0.055
PrsRec=non	2.24	2.25	0.14	0.000	0.000	0.000

of the theme, length difference and person of recipient), the signs of the β-coefficients are the same. This shows that the features have similar effects in both models. The exception is the animacy of the recipient, for which the sign is different in the two models. However, there are indications that both in the traditional model and in the automatic model, the effect is not very stable. First of all, the significance varies across the ten folds: the average p-value is above 0.06, and the standard deviation is above 0.04. Second, we see a significant interaction of animacy with medium in the automatic model, in which the coefficient has the same direction as in the traditional model. Third, in a model that we built on Ice-Auto without any interactions, the animacy of the recipient looses significance completely ($p>0.90$). It also misses significance ($p<0.10$) in a main-effects only model built on Ice-Trad.

The automatic model has five additional significant effects (and a non-significant effect for medium that we kept because of the interactions), presented below the horizontal line. The definiteness of the theme is not significant across the ten folds, but its interaction with medium is. Also significant are the interaction of medium with the animacy of the recipient, the concreteness of the theme, and the pronominality of the recipient. Three of the five additional features thus involve the medium. In Section 4.3 we saw that the FDG parser has more problems with spoken data than with written data (resulting in a much lower precision). Now we see that this has substantially affected the regression model. Apparently, Ice-Auto differs so much from the Ice-Trad that it results in a qualitatively different model.

There are at least three ways to diminish the discrepancy between Ice-Trad

Table 4.9: Significant features in the model built on the 889 instances in ICE-AUTO. Again, the coefficients β for the model fit are provided, together with the average βs in the ten separate models in the ten-fold cross-validation, and their standard deviation. Also, the p-values for the model fit are shown, as well as the average p-values in the ten separate models, with their standard deviation.

Feature	β	av β	st.dev	p	av p	st.dev.
(Intercept)	2.17	2.16	0.16	0.000	0.000	0.000
GivTh=non	-1.55	-1.55	0.13	0.000	0.001	0.000
LenDif	-1.80	-1.81	0.06	0.000	0.000	0.000
AnRec=in	-0.77	-0.77	0.15	0.038	0.063	0.046
PrsRec=non	1.10	1.11	0.08	0.003	0.005	0.003
ConTh=in	-1.46	-1.48	0.09	0.000	0.000	0.000
DefTh=in	0.86	0.86	0.16	0.039	0.062	0.046
PrnRec=p	-1.18	-1.19	0.14	0.000	0.001	0.001
Medium=W	0.16	0.16	0.16	0.711	0.691	0.200
DefTh=in, Medium=W	-1.48	-1.47	0.15	0.006	0.012	0.010
AnRec=in, Medium=W	1.33	1.33	0.20	0.011	0.022	0.019

and ICE-AUTO:[21] (1) by improving the precision of the detection of the cases, (2) by improving the recall of the detection of the cases, and (3) by improving the accuracy of the feature extraction. The second option would mean we either have to use a different parser, or we would have to extend the searches in the FDG parses. We believe this is beyond the scope of this thesis, and we will address this point in our discussion in Section 4.6. The third option seems inefficient, since the accuracies reached by the feature extraction algorithm are already so high that they are surely very difficult to improve (cf. Table 4.6). We therefore choose to improve our data set with the first option: we improve the procedure by inserting a manual step between the detection of the candidates and the feature extraction, in which we manually filter the candidates found.[22] The result is the set of 633 instances, automatically annotated for the features (from now on referred to as ICE-SEMI). There is no significant difference between the proportion of instances from spoken material in ICE-SEMI (60.8%, 385/633) and in ICE-TRAD (60.0%, $\chi^2 = 0.07$, df= 1, $p > 0.75$). This is not true for the proportion of double object constructions: For ICE-SEMI, it is 79.1% (501/633),

[21] In Theijssen, van Halteren, Boves, and Oostdijk (2011a), we show that increasing the data set size is also effective: the prediction accuracy of models applied to data that was annotated automatically is equally good as that found for data with manual annotations, as long as there are enough data points.

[22] All instances in the traditional set have already been checked manually in Chapter 2. In practice, we thus checked only the candidates that were not part of the traditional set.

which is significantly different from the 72.3% in Ice-Trad ($\chi^2 = 9.18$, df= 1, $p < 0.01$). But since the feature effects in logistic regression are very robust against (increasing) class imbalance (Owen, 2007), this will not influence our models.

For Ice-Semi, the model we found was very similar to the traditional model, with one main difference: the concreteness of the theme. It is highly significant in the semi-automatic model, while it did not come even near significance in the traditional model. After all our efforts in developing algorithms to establish concreteness automatically (see also Chapter 3), we thus have to conclude that concreteness is too dependent on the context and on world knowledge to establish it automatically. For this reason, we decided to leave it out, and build a model with 23 features instead, i.e. all features we used before except the concreteness of the theme and its interaction with medium. The resulting model is the model presented in Tables 4.7 and 4.10.

Table 4.10: Significant features in the model built on the 633 instances in Ice-Semi. Again, the coefficients β for the model fit are provided, together with the average βs in the ten separate models in the ten-fold cross-validation, and their standard deviation. Also, the p-values for the model fit are shown, as well as the average p-values in the ten separate models, with their standard deviation.

Feature	β	av β	st.dev	p	av p	st.dev.
(Intercept)	1.85	1.85	0.40	0.018	0.040	0.030
PrnTh=p	1.17	1.18	0.27	0.040	0.070	0.050
GivTh=non	-2.20	-2.22	0.25	0.000	0.000	0.000
LenDif	-2.64	-2.66	0.12	0.000	0.000	0.000
PrsRec=non	1.23	1.24	0.24	0.015	0.028	0.027
PrnRec=p	-1.56	-1.57	0.20	0.001	0.000	0.000

We see that the effects found are indeed very similar to the ones for Ice-Trad in Table 4.8; the correlation between the five βs that are overlapping (those for Intercept, PrnTh, GivTh, LenDif and PrsRec) is 0.97.[23] In comparison with Ice-Auto, the rather unstable effect for the animacy of the recipient has now dropped out of significance, and the pronominality of the theme has become significant. In the comparison between Ice-Trad and Ice-Auto, we saw that the model built on Ice-Auto contained five significant effects more than the model built on Ice-Trad. When comparing the model built on Ice-Semi (Table 4.10) to the one built on Ice-Auto (Table 4.9), we see that three of these have

[23]The β for LenDif was first standardised by multiplying it by the standard deviation of LenDif in the data set.

now neatly disappeared: the interaction of medium with the animacy of the recipient, the definiteness of the theme and its interaction with medium.

The only extra effect that we have in comparison with the ICE-TRAD model is that for the pronominality of the recipient. If we look at the distribution of this feature in the two data sets, we see that the proportion of pronominal recipients is higher in ICE-SEMI (75.7%) than in ICE-TRAD (65.6%). For both data sets, pronominal recipients occur more frequently in double object constructions than in prepositional dative constructions: 88.9% is in a double object construction in ICE-SEMI, and 87.0% in ICE-TRAD. For non-pronominal recipients, there is no clear preference: 48.7% are in a double object constructions in ICE-SEMI, and 44.1% in ICE-TRAD. The pronominality of the recipient thus seems to have a similar distribution with respect to the dative alternation in the two data sets. It only shows up as significant in the ICE-SEMI model because pronominal recipients form a bigger proportion of that set.

We thus conclude that once the low precision of the automatic instance detection is cured, and the concreteness of the theme is left out of consideration, the model is very similar to what we find with a data set that was established completely manually. The semi-automatic model is not really affected by the recall of the detection or the smaller size of the data set. Although we aimed for a completely automatic approach, we have to conclude that human intervention is required, at least when using an off-the-shelf parser like the FDG parser we employed.

4.5.3 Results for the Switchboard data

In this section, we perform an extrinsic evaluation on the test data, the Switchboard data. Given the conclusions of the previous section, we compare the following two models: (1) a model built on the 2,349 instances in SWB-TRAD, and (2) a model built on the semi-automatic set with the 1,292 instances that were also found automatically (SWB-SEMI). In SWB-SEMI, the proportion of double object construction is 83.0% (1,073/1,292), being significantly higher than the proportion of 78.8% in SWB-TRAD ($\chi^2 = 9.43$, df= 1, $p < 0.01$). Again, this is not a problem because logistic regression is robust against class imbalance.

The concreteness of the theme was again excluded from the feature set. The quality of the models is summarised in Table 4.11. Both models show a very good fit to the data, with hardly any overfitting.

The significant effects in the regression models can be found in Tables 4.12 and 4.13. Both show significant effects for the definiteness of the recipient and the theme, the givenness of the theme and the length difference between the theme and the recipient. The coefficients also show the same polarity, and

Table 4.11: Prediction accuracy and concordance *C* for the model fit (*Acc*, *C*) and in 10-fold cross-validation (*av Acc*, *av C*), for SWB-TRAD and SWB-SEMI. The majority baseline and the number of instances are also provided.

Data set	Majority	N	Acc	av Acc	st.dev	*C*	av *C*	st.dev
SWB-TRAD	0.788	2,349	0.933	0.927	0.015	0.972	0.967	0.014
SWB-SEMI	0.830	1,292	0.957	0.954	0.026	0.975	0.969	0.021

their correlation is high (0.97).[24]

Table 4.12: Significant features in the model built on the 2,349 instances in SWB-TRAD. The coefficients β for the model fit are provided, together with the average βs in the ten separate models in the ten-fold cross-validation, and their standard deviation. Also, the *p*-values for the model fit are shown, as well as the average *p*-values in the ten separate models, with their standard deviation.

Feature	β	av β	st.dev	*p*	av *p*	st.dev.
(Intercept)	0.30	0.32	0.14	0.605	0.602	0.153
DefRec=in	0.89	0.89	0.09	0.003	0.007	0.008
DefTh=in	-1.62	-1.63	0.13	0.000	0.000	0.000
PrnRec=p	-0.78	-0.78	0.09	0.008	0.015	0.011
PrnTh=p	1.49	1.48	0.08	0.000	0.000	0.000
GivTh=non	-1.43	-1.43	0.11	0.000	0.000	0.000
GivRec=non	1.31	1.32	0.08	0.000	0.000	0.000
LenDif	-1.61	-1.62	0.07	0.000	0.000	0.000
AnRec=in	1.87	1.88	0.07	0.000	0.000	0.000

The semi-automatic model contains one extra effect: the number of the recipient. As we found for the person of the recipient in the intrinsic evaluation (Section 4.4), this difference seems to be the result of slight differences in the annotation guidelines. Whereas we use a purely syntactic definition, the annotation in Bresnan et al. (2007) is semantic. For instance, when speaking about a hypothetical person, speakers sometimes switch to plural *them* to refer to such persons. We label *them* as plural, while the annotations in SWB-TRAD call it singular. This was the case in 38 of the 64 disagreeing annotations. For 14 more, the disagreement was caused by a different treatment of the noun *people*, being semantically plural (*a group of persons*), but syntactically singular

[24] Again, we standardised length difference by multiplying the β by the standard deviation of the feature in the data.

Table 4.13: Significant features in the model built on the 1,292 instances in Swb-Semi. Again, the coefficients β for the model fit are provided, together with the average βs in the ten separate models in the ten-fold cross-validation, and their standard deviation. Also, the p-values for the model fit are shown, as well as the average p-values in the ten separate models, with their standard deviation.

Feature	β	av β	st.dev	p	av p	st.dev.
(Intercept)	1.54	1.57	0.15	0.009	0.012	0.008
DefRec=in	3.55	3.56	0.20	0.000	0.000	0.000
DefTh=in	-2.09	-2.10	0.20	0.000	0.001	0.002
GivTh=non	-1.58	-1.59	0.20	0.004	0.007	0.005
LenDif	-3.60	-3.62	0.16	0.000	0.000	0.000
NrRec=sg	-0.93	-0.93	0.06	0.003	0.005	0.002

(plural: *peoples*).

The traditional model contains four more significant effects that were not found in the semi-automatic model: the pronominality of the theme, the animacy of the recipient, the givenness of the recipient and the pronominality of the recipient. For the latter three, the problem is that they are correlated: all three very frequently have the same value. This is because many recipients consist of personal pronouns only, and they are always pronominal, definite and discourse given, and animate most of the time. Because these features are correlated, it is not surprising that not all of them show up in the semi-automatic model, which is based on fewer data points than the traditional model.

As with the models for the ICE data, we see some differences between the models built on Swb-Trad and on Swb-Semi. But these differences do not seem to be caused by the quality of the automatic approach, but by difficulties in the data itself: the use of different annotation definitions and the correlation of many of the features. The low recall of 55.0% may explain the lack of significance for some of those correlated features.

4.6 Discussion

Besides the difficulty of collecting a suitable data set that can be used to model variation in language (e.g. syntactic alternation), linguists taking such a modelling approach have a more fundamental challenge to meet. When using modelling techniques such as logistic regression, one models the data that is offered. Two different data sets, though drawn from the same population, can result in different models because their composition differs. Because we use

two different samples (traditional and automatic) from the English language as represented in the ICE-GB and Switchboard corpora, this accidental composition could affect the models. It is not clear whether the traditional set is closer to the actual English language than the automatic set. The models found for either of the data sets are not necessarily true, and the features that show no significance in our models could still play a role in another data set. Moreover, there is still no consensus about the definitions of the features we have employed. The definitions we used for this chapter are chosen such that they allow comparison with previous work (Chapter 2, Bresnan et al., 2007), but they are by no means definitive. Moreover, we have seen that even the definitions of which we believed they were the same, appeared to be slightly different after all.

The effect of the composition of a data set usually grows when data sets become smaller. In the near future, we will therefore apply the procedure to a larger corpus: the one-hundred-million-word British National Corpus (BNC Consortium, 2007). The results found for this data set may show whether the almost significant effects turn up really significant when larger amounts of data are considered. A possible drawback is that we have shown that a fully automatic approach is not accurate enough. Instead, we need a semi-automatic approach in which we manually filter the candidates suggested by the parser. For a large corpus as the BNC, this step may take considerable time. However, some preliminary annotation work shows that the manual checking is not as time-consuming as one would think: with the help of a user-friendly interface, one can check up to 200 candidates per hour. Moreover, the inter-annotator agreement for this task is comforting: an average κ of 0.74 (for four annotators who all checked the same 100 candidates).

One might wonder if the human intervention is still needed when employing a different parser. In this thesis, we have decided to use an off-the-shelf syntactic parser that distinguishes both dative constructions explicitly. Parsers that have this information available are rare, and we believe human intervention will always be necessary. Of course, such a manual step, in which one checks the candidates suggested by the parser, can also be performed on the output of other parsers that may or may not recognise dative constructions explicitly. One could for instance decide to employ a parser that does not distinguish between prepositional dative constructions and locative constructions (e.g. *I brought him to school*), but that yields a higher recall. Another possibility is to improve an existing parser by training on data that is similar to the data studied. However, this is a difficult procedure that requires one to have experience with parsing. Our approach has shown that even with an off-the-shelf parser, that yields a low recall, sensible data sets can be obtained. This is a promising result for corpus linguists who study a syntactic phenomenon but do

not have access to syntactically annotated data.

In fact, the semi-automatic approach is also suitable for research on different syntactic alternations. One could select a parser that seems to perform well on the construction in question, and then manually check the proposed candidates. When seeing recurring patterns, one can add simple heuristic rules like we formulated for the dative alternation. Next, one can use the feature extraction script presented in this chapter.[25] Many of the features included in the script are generally known to be relevant for other syntactic alternations, as already noted in Section 4.1. The script should be provided with three bits of information for each noun phrase that needs annotation: (1) which word is the syntactic head, (2) what are the lemmas of the words in the noun phrase, and (3) what are the POS tags of these words. Using a different parser would thus mean that the extraction script needs some adjustments. For establishing the discourse givenness, it also needs to have the preceding context. The selection of a corpus thus also leads to the need for some minor changes in the extraction script, so it can deal with the corpus input provided.

4.7 Conclusion

In this chapter, we have addressed the question of whether automatically obtained and enriched data is suitable for use in linguistic research on syntactic alternations, even if the data may contain errors. We have taken the English dative alternation as a case study. This offered us a way to evaluate the automatically obtained data extrinsically, namely by employing it to build logistic regression models like those in Bresnan et al. (2007). We employed two data sets that were manually obtained: 930 instances collected in Chapter 2 from the ICE-GB corpus of spoken and written British English (Ice-Trad), and 2,349 instances collected by Bresnan et al. (2007) from the Switchboard corpus of spoken American English (Swb-Trad). The first data set has been employed to tailor the automatic approaches, and to evaluate the errors made. The second data set has not been seen previously, and has been used as a test set in quantitative evaluations. With respect to the aforementioned question, there are two main conclusions to be drawn.

First, we have to conclude that the FDG parser that we employed is not very successful in detecting instances of the dative alternation. In combination with our filtering heuristics, the recall was 66.6% for the instances found automatically in the ICE-GB (Ice-Auto) and 55.0% for those found in Switchboard (Swb-Auto). For precision, we reached 69.6% for Ice-Auto and only 48.0% for

[25] The feature extraction script can be downloaded from `http://daphnetheijssen.ruhosting.nl/downloads`.

Swb-Auto. The analysis of the errors in Ice-Auto showed that the FDG parser has most difficulty with spoken material, with longer sentences and with PP-attachment. Parse errors were the main cause of missing instances (decreasing recall) and incorrectly accepting candidates (decreasing precision). Seeing the nature of the Switchboard data (spontaneous speech only, with many disfluencies), it is not surprising that the FDG parser has great difficulty recognising dative constructions. The regression model for Ice-Auto contained four significant effects that were not found for Ice-Trad. We concluded that Ice-Auto contained too many errors to give the same – or at least similar – results as those obtained for Ice-Trad. We solved this problem by inserting one (simple) manual step: manually checking the relevance of the candidates that were found automatically, before annotating the approved instances automatically. The model built on only the 633 instances that were manually approved (Ice-Semi) appeared to be very similar to the one found for Ice-Trad. This is also what we found for the 1,292 approved candidates in the semi-automatic Switchboard set (Swb-Semi).

Second, we conclude that our rather straightforward feature extraction algorithm is suitable for automatically annotating the instances with information that is syntactic (e.g. number), semantic (e.g. animacy) and discourse-related (e.g. givenness) in nature. The κ scores between the manual and the automatic annotations were similar to scores found between human annotators, except for the intuitively most difficult features: animacy, concreteness and discourse givenness. Only the automatic annotation of the concreteness of theme was so dissimilar from the human annotations that it notably influenced the regression models. When excluding this feature, the models built on Ice-Semi and Swb-Semi (with the automatic annotations) were very similar to the ones obtained for Ice-Trad and Swb-Trad (with manual annotations). The differences we found did not seem to be caused by the errors in the automatic annotations, but by properties inherent to the data sets: multiple correlations between the features, and the presence of different definitions for the same feature.

In sum, we see that the models found for the automatic data sets are especially hampered by the presence of candidates that are not really instances of the dative alternation, but that were included due to errors in the automatic analyses. We also have to conclude that establishing the concreteness of nouns automatically is a bridge too far. But when the instances found are manually checked for relevance, and concreteness is left out of consideration, the models found are very similar to the ones found for traditionally established data sets.

5

Comparison of speaker groups

Edited from: Theijssen, D., Bresnan, J., Ford, M., & Boves, L. (2011). *In a land far far away... A probabilistic account of the dative alternation in British, American, and Australian English.* (Under review.)[1]

Abstract

This chapter presents a corpus and judgement study of the dative alternation, which are performed in a framework of probabilistic linguistics in which we assume that syntactic structure is influenced by linguistic factors whose relative importance may vary. With regression models, we compared the dative alternation of British, American, and Australian speakers of English varying in age and gender. We found that both in produced speech and in judgements, the linguistic factors show a consistent pattern (*harmonic alignment*) across different varieties, age groups, and genders: animate objects usually precede inanimate objects, definites precede indefinites, shorter precede longer, and pronouns precede nonpronouns. The two studies also revealed subtle distributional differences between the roles that these linguistic factors play across the different speaker groups.

[1]This study was supported by the Fulbright Center (personal grant to Daphne Theijssen), and by the National Science Foundation (grant numbers IIS-0624345 and BCS-2=1025602 to Stanford University, PI Joan Bresnan).

5.1 Introduction

The theoretical framework in which we study the dative alternation in this chapter is that of probabilistic linguistics. We build on existing research on the dative alternation, making use of the predictive features that have been suggested in previous research. In our interpretation of the probabilistic linguistics framework, we assume that these features are suitable for the study of syntactic structure, and we investigate their relative contribution to the likelihood of the two constructions. These assumptions are inspired by the line of research initiated by Bresnan et al. (2007).

As mentioned in Chapter 1, parallels to the English dative alternation also occur in other languages, e.g. in Dutch (Colleman, 2006), Korean (Choi, 2007), Japanese (Miyagawa & Tsujioka, 2004), Greek (Anagnostopoulou, 2005) and Spanish (Beavers & Nishida, 2010). The study of alternations across languages is very useful for gaining knowledge about syntactic variation (e.g. Levin, 2008). However, cross-linguistic studies are not straightforward, because there are many differences between languages that can influence the variation. We therefore choose to study alternation in different variants of the same language: English. Over time, the different varieties of English have evolved in different ways, slowly altering the probabilistic distributions of noun phrases and syntactic constructions in each variety. By studying the dative alternation in different varieties of English, we aim at a (fine-grained) description of the differences in the likelihood of the two syntactic constructions across the varieties, and of the influence of various features on this likelihood.

A number of researchers have already investigated similarities and differences between the dative alternation in different varieties of English. Examples are comparisons of the dative alternation in American and British writing (Grimm & Bresnan, 2009), in various written genres in American and British English over time (Wolk et al., 2012), in speech by people from New Zealand and the USA (Bresnan & Hay, 2008), in various experiments conducted with American and Australian participants (Bresnan & Ford, 2010), in speech and writing in Indian and British English (Mukherjee & Hoffmann, 2006), and speech and writing in American and African American English (Kendall, Bresnan, & van Herk, 2011). Overall, it seems that the same features play a role in the dative alternation across different varieties of English, but that their relative importance may vary.

Many comparisons between the dative alternation in different varieties of English are still to be made, even in the varieties that have already been included in previous research. For instance, there are several studies that include British or Australian English (e.g. Mukherjee & Hoffmann, 2006; Grimm & Bresnan, 2009; Bresnan & Ford, 2010), but there is no direct comparison

between the two varieties. The present chapter provides multivariate studies of the dative alternation in British English, American English and Australian English. Our first research question is thus: What are the differences and/or similarities in the dative alternation in *British, American and Australian English*?

Given the findings in existing work, we expect to find the same general patterns in the roles that the features play in the three varieties, but with subtle distributional differences. British English was the origin of both American and Australian English, but at different times in history. American English came into existence after the arrival of the British in the early seventeenth century, while Australia was colonised much later, in the late eighteenth century. We therefore could expect to find other differences between American and British English, than between Australian and British English. On the other hand, the ubiquitous cultural influences that are typical of our present society, e.g. the wide-spread influence of American culture, may now be increasing the similarity of the three varieties.

It is now widely known that language is not only influenced by linguistic, but also by *extralinguistic* factors (e.g. Gregory, 1967; Biber, 1985). Recently, some researchers have extended existing multivariate models, based on linguistic features, by adding extralinguistic factors that may facilitate language variation and change (for an overview see Kristiansen & Dirven, 2008; Geeraerts et al., 2010). Language variety, the topic of our first research question, can be considered an extralinguistic factor. But there are also examples of extralinguistic factors that are nested in a language variety: properties of the text such as genre and modality, and speaker characteristics such as age and gender.

The effect that extralinguistic factors (other than language variety) have on the dative alternation has received only little attention in multivariate analyses. Much more attention has been paid to this in studies on the *genitive* alternation: *John's book* vs. *the book of John* (e.g. Hinrichs & Szmrecsányi, 2007; Szmrecsányi & Hinrichs, 2008; Tagliamonte & Jarmasz, 2008; Jankowski, 2009; Grafmiller, 2012). Szmrecsányi (2010) shows that the linguistic factors play a role across time, varieties, modalities and genres, while the differences in the exact roles of these factors can be explained by extralinguistic factors.

For the *dative* alternation it has been found that in New Zealand English, young and elderly speakers favour the prepositional construction more than middle-aged speakers (Bresnan & Hay, 2008). Also, Bresnan and Ford (2010) found a near-significant trend for male Australians to produce more prepositional dative constructions than female Australians. In this chapter, we study the effect of the sociolinguistic factors *age* and *gender* on the dative alternation in the aforementioned three varieties English. Our second research question

is thus an extension of the first: What are the differences and/or similarities in the dative alternation of British, American and Australian language users varying in *age and gender*? Given the finding of Bresnan and Ford (2010), we expect male Australians to be more positive towards the prepositional dative construction than female Australians.

The goal of this chapter is to answer the two aforementioned research questions, one of which addresses the dative alternation in three varieties of English, and one the effect of age and gender on the dative alternation in these varieties. We perform two studies to reach this goal, placed in a probabilistic framework and making use of multivariate models.

Our first study is a corpus study. In corpus-driven research, theories are based on the *results* of language production that have been collected in natural settings, where speakers and writers are usually unaware of their language behaviour. We include only utterances made in spontaneous (unscripted) speech, since these are the most natural instances of language. In our corpus study, we compare the dative alternation in spontaneous speech in British and American English.[2]

It is rather difficult to investigate the effect of age and gender in corpus data, because (1) corpora often lack sufficient meta-data to provide the information needed, and (2) despite the common use of datives in English, it is difficult to find enough occurrences to investigate the variables, especially since so many other features influence the choice between the two syntactic constructions. For these reasons, we perform a second study: a judgement study in which participants rate the naturalness of the two constructions in context. In fact, the study is an extension of that in Bresnan and Ford (2010), but using a web-based version of the original judgement study on paper, and including participants from the US, Australia and the UK, with varying ages. Bresnan and Ford (2010, p. 201) found that the participants have 'strong predictive capacities, preferring and anticipating the more probable of two alternative syntactic paraphrases'. Thus, the judgements of the participants closely resembled the probabilities found in their corpus study. This is in line with the recent development of models in which processes used in production are directly linked to the processes used in comprehension. Pickering and Garrod (2005) suggest that people use language production processes to make predictions about the language they hear during comprehension. For this reason, distributional differences found in language production (corpus data) should also be found in perception and comprehension experiments. This motivates our use of a judgement study.

[2]Australian English is not included due to our lack of a comparable Australian corpus (containing spontaneous speech collected in the early 1990s).

The remainder of this chapter is structured as follows: Section 5.2 presents existing work on the dative alternation in English. In Section 5.3, we present our corpus study; the judgement study is the topic of Section 5.4. Our discussion and conclusion is provided in Section 5.5.

5.2 Related work

In Chapter 1, we already mentioned that Bresnan et al. (2007) used fourteen features in multivariate regression models for a data set extracted from the Switchboard corpus of spoken American English (Godfrey et al., 1992). The same features were employed in Chapter 2, being a study of the dative alternation in the ICE-GB corpus (Greenbaum, 1996), containing spoken and written British English. The features mostly describe characteristics of the theme (*the poisonous apple* in Chapter 1) and the recipient (*Snow White*). In both studies, the same pattern appeared. Everything else being equal:

animate usually precedes inanimate
definite usually precedes indefinite
given usually precedes nongiven
shorter usually precedes longer
pronoun usually precedes nonpronoun

This consistent pattern is sometimes referred to as *harmonic alignment*, and it has been found in several other varieties of English, e.g. in New Zealand English (Bresnan & Hay, 2008), Australian English (Bresnan & Ford, 2010) and African American English (Kendall et al., 2011).

Although the harmonic alignment is consistent across different varieties of English, several studies have found distributional differences in the role that the features play. In a study of American and British writing from the 1960s and 1990s, Grimm and Bresnan (2009) found that in the 1990s British writers were more likely to use a personal pronoun as the second object of a double-object construction (e.g. *give the man it*) than American writers. The British flexibility with respect to pronominality can also be found in the fact that some British dialects allow reversed double object constructions such as *give it him* (Siewierska & Hollmann, 2007; Haddican, 2010). With respect to changes over time, Grimm and Bresnan (2009) found that both British and American English showed an increasing tendency to use the double object construction. The American data showed that the effect of pronominality was stronger in the 1990s than in the 1960s. For the British data, the effect is the opposite: the effects of pronominality and thematicity (an approximation of discourse givenness) were stronger in the 1960s than in the 1990s.

Bresnan and Ford (2010) used psycholinguistic experiments and found differences in the effects of object length between American and Australian English. In a judgement study, they found that as the recipient increased in length relative to the theme, the Australian participants showed a greater preference for the prepositional dative than the US participants. In a task measuring reaction times while reading datives, they found that as the theme increased in length in prepositional datives, the US participants showed a more rapid (steeper) slowing down in reaction time than the Australians. Bresnan and Ford suggested that the Australians might have a stronger preference for the prepositional dative compared to US participants.

In the diachronic corpus study by Wolk et al. (2012), based on the ARCHER corpus (A Representative Corpus of Historical English Registers, Biber, Finegan, & Atkinson, 1994), the effect of theme length (measured in characters) was stronger in American than in British English writing. There were also some changes over time independent of language variety: the double object construction has become more popular, the effects of the animacy and the length of the recipient have *decreased*, and the effects of the pronominality and definiteness of the recipient have *increased*. Wolk et al. also observe that the more oral registers tend to favour the double object construction, while more literate registers contain relatively more occurrences of prepositional dative constructions. The same has also been observed by Bresnan et al. (2007) in their comparison of the dative alternation in spoken telephone dialogues (Switchboard) and written news paper texts (Wall Street Journal texts in the Penn Treebank, Marcus, Marcinkiewicz, & Santorini, 1993).

Bresnan and Hay (2008) found a stronger effect of animacy in spoken New Zealand English than in spoken American English. They also found that young and elderly New Zealanders favoured the prepositional construction more than middle-aged speakers. The corpus study by Mukherjee and Hoffmann (2006) showed that the prepositional dative is more frequent in the Indian than in the British English components of the International Corpus of English, ICE. The corpus study by Kendall et al. (2011) revealed no differences between the dative alternation in American and African American English speech and writing.

5.3 Speech corpus study: British and American English

Previous studies that compare the dative alternation in the two related varieties British and American English have shown that there are many similarities, but also some subtle distributional differences. This is particularly interesting because American English originated from British English in the early sev-

enteenth century, but has evolved into a clearly different variant of English, embedded in its own nation and culture.

The existing work has focussed on written corpus data. Although written data can be very informative (especially when spoken data is not available, as for the historical study in Wolk et al., 2012), there is no question that spontaneous speech is the most natural language form. Therefore, we compare the dative alternation in British and American English in spontaneous speech corpus data. Our study will shed new light on the role that the linguistic features play in the dative alternation in two fairly different, but closely related, varieties of English. For this reason, it will provide us with new evidence for the universality of the features and their distributional differences in different varieties.

5.3.1 Data

Our American corpus data is a corrected version of the data described in Bresnan et al. (2007). It consists of 2,349 instances taken from the Switchboard corpus of American telephone dialogues, collected in the early 1990s. All instances were manually checked for relevance and manually annotated with the features introduced in Table 1.1 in Chapter 1. In the current chapter, we only employ the seven strongest and least correlated features, which are presented in Table 5.1. For details about the data, refer to Bresnan and Ford (2010). There are 38 different verb types.

Table 5.1: Features adapted from Bresnan et al. (2007). Only the features that are strongest and least correlated with each other are included in the present chapter, using the values provided.

Feature	Description	Values (shortened)
AnRec	Animacy of the recipient	animate (a), inanimate (in)
ConTh	Concreteness of theme	concrete (c), inconcrete (in)
DefRec	Definiteness of recipient	definite (d), indefinite (in)
DefTh	Definiteness of theme	definite (d), indefinite (in)
LenDif	Length difference (log of ratio)	ln(words th)−ln(words rec)
PrnRec	Pronominality of recipient	pronoun (p), nonpronoun (n)
PrnTh	Pronominality of theme	pronoun (p), nonpronoun (n)

The British corpus data is that of Chapter 2 and consists of 930 instances from the British component of ICE. This corpus contains written and spoken English in various genres. It is annotated for the same features as the American corpus data (cf. Chapter 2 for details). We limit ourselves to the 491 instances of

unscripted speech, containing 41 different verb types. The spoken part of the ICE-GB Corpus was collected in 1990-1992, the same period as Switchboard. It contains all kinds of spontaneous speech, including face-to-face spoken dialogues, but no telephone dialogues.

Given the different verb types included in the two data sets, we decided to remove all instances with verbs present in only one of the two sets. We thus removed from the American data the 264 instances in which the verb was *afford*, *allot*, *allow*, *award*, *bet*, *cost*, *deny*, *flip*, *float*, *loan*, *mail*, *promise*, *serve*, *swap*, *take* and *wish*. From the British data we removed the 35 instances with *bowl*, *circulate*, *deliver*, *explain*, *get*, *guarantee*, *let*, *mouth*, *open*, *pass*, *play*, *pose*, *present*, *report*, *return*, *sign*, *spread*, *square* and *suggest*. The resulting data set consists of 2,541 instances with 22 different verb types (see Table 5.2).

Table 5.2: Verb types and frequencies in the combined corpus data

Verb	freq	Verb	freq	Verb	freq	Verb	freq
give	1,512	teach	65	write	19	cause	12
send	170	bring	46	feed	16	read	11
tell	168	offer	46	leave	15	assign	5
pay	143	charge	42	lend	15	make	5
show	94	do	38	hand	13	quote	3
sell	79	owe	24				

The American and British corpus data were created separately, each with their own (single) annotator. To establish the similarity between the annotations in the two sets, the annotator of the British set (the first author) annotated 30 items in the American set. The κ scores for the aforementioned features were all ≥ 0.78, showing good overall agreement. We included the feature for the length difference between the theme and the recipient (LenDif), following the definition in Chapter 2, as presented in Table 5.1.

5.3.2 Method

We compare the two varieties by including language Variety as a fixed main factor, as well as its interactions with the other features, in a mixed logistic regression model.[3] The features are potentially correlated; for instance, pronominal objects are generally shorter than full noun phrases because they often consist of a single pronoun only. The point-biserial correlation coefficients confirm this: 0.40 between LenDif and PrnTh, and 0.42 between LenDif

[3]We used `lmer()` in the `lme4` package in `R` (R Development Core Team, 2008).

and PrnRec. We therefore residualise LenDif: we include PrnRec and PrnTh in a linear regression model that predicts LenDif. The unexplained variance (the residuals) is then included as a fixed factor (rLenDif) in the eventual logistic regression model, replacing LenDif.

The lemma of the verb (Verb) is included as a random factor, as well as the pair of the lemma of the verb and the lemma of the head of the theme (VerbThHead). This is to capture strong biases in certain expressions, for instance 'pay attention to someone' (VerbThHead = *pay_attention*), and 'give someone the creeps' (VerbThHead = *give_creep*). One could decide to exclude such instances, since they do not seem to allow alternation at first sight (cf. Ozón, 2009, who excluded instances like 'give it a chance'). However, Google searches indicate that alternation often is possible in these expressions (cf. Bresnan et al., 2007), so we decided to retain them. Since many of these VerbThHeads are infrequent, we group all that have a frequency < 3 in a category 'other'.

The variable selection approach we take is the following: We first include all fixed factors and all interactions with Variety, and remove all *interactions* that are not significant, in a single step. Next, we continue with stepwise backward elimination, removing the most insignificant feature until no insignificant ones remain. If an interaction is significant, we always keep the features it consists of as main features as well. To test the final model for overfitting, we employ 10-fold cross-validation: A regression model is fitted to a subset of 90% of the data. The random effect values (the best linear unbiased predictors, or BLUPs) and coefficients (for the features, or 'fixed factors') are then used to predict the remaining 10% of the data. This is repeated ten times so that the whole data has served as a test set. The division in test sets is random. If a test set contains a Verb or VerbThHead that is not part of the training set, the random effect value is set to that of the category 'other'. The number of items per test set is 254, of which at least 36 and at most 55 are British. We present the average of the concordance C, the coefficients and the p-values across the ten test sets, together with their standard deviations.

5.3.3 Results and discussion

Table 5.3 presents information about the regression model we found. The c-number[4] for the residualised variables was 10.34 before variable selection, which indicates that there is mild collinearity. The concordance (C index)[5] is above 0.97, also in 10-fold cross-validation, which shows very good overall fit.

[4]We used `collin.fnc()` in the `languageR` package in `R`.
[5]We used `somers2()` in the `Hmisc` package in `R`.

Table 5.3: Characteristics of the regression model for the combined spoken British and American English corpus data (1 = prepositional dative)

Description	Value
Collinearity (c-number) before variable selection	10.34
Concordance (C index) of model fitted on all data	0.984
Average concordance (10-fold cv)	0.972
Standard deviation concordance (10-fold cv)	0.009
Number of observations	2,541

In regression, the random factors are treated as normal distributions with mean 0. Table 5.4 provides the standard deviations of these distributions. Both Verb and VerbThHead are significant, even if both are included at the same time ($p < 0.001$, using ANOVAs on models with the same fixed factors and nested random effects). In fact, most of the variance is explained by the random effects, which alone (i.e. without any of the features) yield a concordance of 0.932.

Table 5.4: Number of values and standard deviations of the random effects in the regression model fitted on all corpus data (1 = prepositional dative)

Random effect	nr of values	stdev.
VerbThHead	189	2.38
Verb	22	1.88

The coefficients of the features (the main effects), together with their p-values, are shown in Table 5.5. Positive coefficients indicate that the feature value in the first column increases the likelihood that the construction used is the prepositional dative, negative ones increase the likelihood for the double object construction. Features that are not significant, but kept in because they are part of a significant interaction, are written in parentheses. All significant main effects are in line with the *harmonic alignment* explained in Section 5.2: pronominal themes and inanimate and indefinite recipients favour the prepositional dative construction, longer themes and pronominal recipients the double object construction.

From the interaction term *DefTh=in:Variety=US*, we see that American speakers show a significantly stronger tendency towards the double object construction if the theme is indefinite (e.g. *give him a book*), as compared to British speakers. There is also an interaction between variety and length difference (*rLenDif:Variety=US*). While there is a main effect of length differ-

Table 5.5: Coefficients and their *p*-values for the combined spoken British and American English corpus data, fitted on all corpus data (1 = prepositional dative). The average and standard deviation in 10-fold cross-validation are also provided (the models fitted on 90% evaluated on the remaining 10%). Insignificant features are in parentheses

Feature	coef	average	stdev	*p*-value	average	stdev
(Intercept)	-0.35	-0.37	0.23	0.620	0.627	0.213
AnRec=in	1.42	1.43	0.11	0.000	0.000	0.000
DefRec=in	1.42	1.43	0.13	0.000	0.000	0.000
(DefTh=in)	-0.56	-0.55	0.17	0.318	0.371	0.158
PrnRec=p	-4.12	-4.15	0.10	0.000	0.000	0.000
PrnTh=p	2.83	2.86	0.15	0.000	0.000	0.000
rLenDif	-2.58	-2.60	0.09	0.000	0.000	0.000
(Variety=US)	0.67	0.69	0.14	0.150	0.172	0.081
DefTh=in:Variety=US	-1.34	-1.39	0.17	0.026	0.036	0.029
rLenDif:Variety=US	0.93	0.93	0.12	0.027	0.039	0.018

ence, such that there is a greater preference for the double object dative as the theme increases in length compared to the recipient, this effect is less strong for the American speakers.

In their diachronic corpus study, Wolk et al. (2012) found that in both British and American English writing, the effect of the length of the recipient has decreased, and the effect of the definiteness of the recipient has increased. In our study of spoken data, we found that the definiteness of the theme is more important, and the length difference is less important in American English than in British English. It thus seems that, in comparison with the diachronic study in Wolk et al. (2012), American English has evolved even further than British English, with respect to definiteness and length. However, there are many differences between the study by Wolk et al. (2012) and our study, most notably the medium of data (spoken versus written) and the type of data used (diachronic and modern English). For this reason, the fact that our results seem to be in line with the findings in Wolk et al. et al. should be tested further.

Overall, our corpus study has shown that in contemporary spontaneous spoken English, most of the features suggested in the literature play similar roles in British and American English. These roles are in line with *harmonic alignment*. However, it seems that the definiteness of the theme plays a significant role in American English, but not in British English. With respect to length difference, we saw that it is relevant across British and American English, but that its effect size differs. This provides novel evidence that there are slightly different distributional patterns in the production of two varieties of English.

5.4 Judgement study: Age and gender in British, American, and Australian English

In this section, we investigate how natural the same dative sentences seem in context, according to speakers of British, American, and Australian English. The participants vary in gender and age, which enables us to answer our second research question: What are the differences/similarities in the judgement of dative sentences in British, American, and Australian English made by participants varying in age and gender?

Effects of the *linguistic* factors used throughout this thesis have been found in various experimental studies (e.g. Bock & Irwin, 1980; Bock et al., 1992; Arnold et al., 2000; Prat-Sala & Branigan, 2000; Rosenbach, 2005; Bresnan & Hay, 2008; Bresnan & Ford, 2010). People seem to use language production processes to make predictions in language comprehension (Pickering & Garrod, 2005). Therefore, a judgement study will provide insight in the effect of age and gender on the dative alternation in the three varieties.

Until now, the effects of age and gender have been mostly ignored in research on the dative alternation in English. Exceptions are Bresnan and Hay (2008), who found that in New Zealand English, young and elderly speakers favour the prepositional construction more than middle-aged speakers, and Bresnan and Ford (2010), who found a (near-significant) tendency for male Australians to be more likely to produce a prepositional dative construction than female Australians. Since so little is known about the effect of age and gender on the dative alternation in English, we provide such a study for British, American, and Australian English.

5.4.1 Experimental setup

We extend the judgement study of Bresnan and Ford (2010) by including British English, using a wider age range (20 to 65 years), and conducting it through a website instead of on paper. Participants had to read a short passage followed by two possible continuations: one with a double object construction, one with a prepositional dative. They were asked to rate the naturalness of both options by dividing 100 points between them: the more points, the more natural. A screenshot of the experiment website is provided in Figure 5.1. All items were presented in random order, and the order in which the two options were presented was alternated.

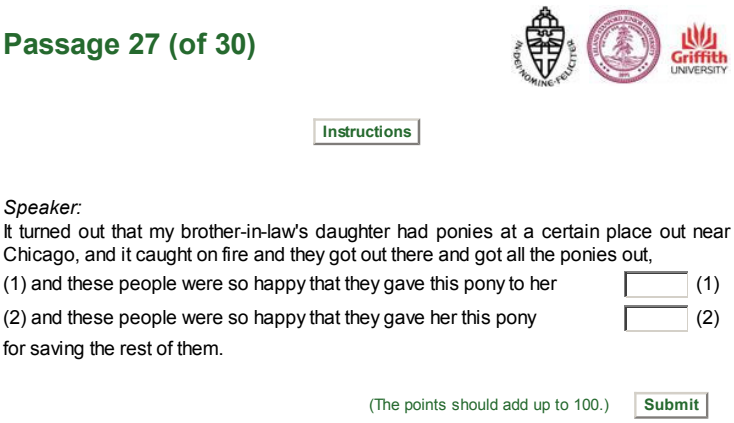

Figure 5.1: Screenshot of the experiment website

5.4.2 Items and participants

We use the same 30 items taken from the Switchboard corpus as in Bresnan and Ford (2010), who already localised these for Australian English by replacing US-specific vocabulary and place names; we did the same for British English.[6] The verbs, themes, and recipients in the dative constructions were not altered, thus keeping the items comparable across the three varieties.

The participants were all volunteers who were entered in a prize draw for a gift voucher. Table 5.6 shows the characteristics of the participants.

Table 5.6: Characteristics of British (UK), American (US) and Australian (Aus) participants in the judgement study

	Female					Male				
	N	av age	stdev	min	max	N	av age	stdev	min	max
UK	22	32.0	11.7	21	61	18	31.5	10.4	21	63
US	22	37.3	14.3	21	65	13	32.1	11.7	21	61
Aus	23	35.3	12.3	23	63	17	32.1	11.8	20	64

[6] We thank Dr. Caroline Piercy for localising the items for British English.

5.4.3 Modelling

Using linear regression, we modelled the participants' ratings for the prepositional dative variant, being a number between 0 and 100. The features we included as fixed factors are the same as those discussed in Section 5.3, except that we now residualise length difference on the definiteness and the pronominality of the recipient and the theme. The reason for this is that for the 30 items used, there is not only a high correlation of length difference with the two pronominality features (0.80 for the recipient, and 0.49 for the theme), but also with definiteness (0.29 for the recipient, and 0.32 for the theme).

Besides these main factors, we also look at Age, Gender, and Variety, and their interaction with the other features. We also control for the random order in which the items were presented (ItemOrder, ordinal with values 1-30), the alternating order of the two options (OptionsOrder, binary) and the rating assigned to the previous item (PrevRating, interval scale) by including these as fixed factors.

The verb lemma (Verb) is included as a random factor. Some participants generally award higher scores to the prepositional dative variant than other participants. Also, some participants use the whole range of points (0-100), while others use only a portion of it (e.g. 30-70). With our regression models, we want to establish the relative influence of the explanatory features between participant groups (the three nationalities, age and gender), not within these groups. We therefore follow the approach in Bresnan and Ford (2010) and correct the individual participant differences by including a random intercept for participant (Participant) and a random slope of participant over the centred predicted corpus probabilities (cCorpusProbs). The predicted corpus probabilities are taken from a logistic regression model built on the 2,349 spoken American instances, including the verb theme head as a random intercept. The predicted corpus probabilities are centred by subtracting the mean, as suggested in the literature (e.g. Baayen, 2008).

We apply the same variable selection method as in Section 5.3. We first build three separate regression models for the three varieties of English, including two-way interactions between the features and Age and Gender. Next, we build a single model for all three varieties, only including the interactions that showed up as significant in at least one of the three separate models. The models are evaluated by R^2. We again also do this in a 10-fold cross-validation setting: we randomly divide the 3,450 items in 10 test sets with 345 items each. On each test set, we apply the regression model fitted on the remaining 3,105 items, and calculate the R^2. We also establish the average and standard deviation of the coefficients and t-values[7] in the ten models.

[7] P-values are not included because there is still uncertainty in the field on how to establish

5.4.4 Results for the individual models

Figure 5.2 presents the significant coefficients in the individual models found for the ratings by participants of the same variety of English. The values for PrevRating and Age have been divided by 10 to increase the coefficients with a factor 10 (to make them more visible in the Figure).

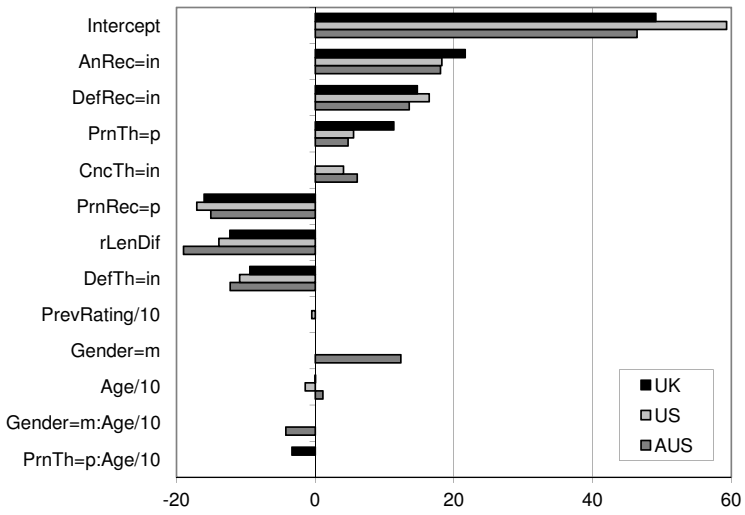

Figure 5.2: Coefficients in the final models, made separately for the British, American, and Australian participants. Positive coefficients favour the prepositional dative construction, negative coefficients the double object construction.

The random effects for Verb, Participant and Participant over cCorpusProbs are significant, also in combination with each other ($p < 0.001$, using ANOVAs on models with the same fixed factors and nested random effects). The R^2 is 0.514 for the UK model, 0.499 for the US model and 0.584 for the Aus model, which means that the models are able to explain about half of the variance. This is similar to what was found in the study we extended ($R^2 = 0.529$, Bresnan & Ford, 2010). A large part of the variance is explained by the random effects: the random-effects-only models reach R^2s of 0.475 (UK), 0.457 (US) and 0.505 (Aus).

Overall, the effects of the features are again in accordance with *harmonic alignment*. The bar for *PrnTh=p:Age/10* in Figure 5.2 shows that older British participants assign lower ratings to the prepositional dative construction when

the number of degrees of freedom in mixed-effect linear regression models. We consider significant the features with an absolute t-value > 2 (Baayen, 2008).

the theme is a pronoun than younger British participants. Younger British people thus seem to have a stronger preference for structures such as *I gave it to the man* (not *gave the man it*) than older people.

Figure 5.2 also shows that in the judgements made by the American participants, older American participants give lower ratings for the prepositional dative than younger participants (indicated by the bar for *Age/10*). This seems to indicate that in the US, the prepositional dative construction is becoming slightly more popular.

Male Australian participants give higher ratings to the prepositional dative than female Australian participants (the bar for *Gender=m* in Figure 5.2), which supports the findings in Bresnan and Ford (2010). The effect is less strong for older than for younger Australian males (because of the significant interaction *Gender=m:Age/10*). We could thus say that younger Australian men seem to be more positive towards using the prepositional dative construction.

5.4.5 Results for the combined model

All features in Figure 5.2 are next included in the combined model, together with their interaction with language Variety. The characteristics of the final model can be found in Table 5.7, the standard deviations of the random effects in Table 5.8.

Table 5.7: Characteristics of the regression model for the ratings for the prepositional dative construction by British, American, and Australian participants (combined)

Description	Value
Collinearity (*c*-number) before variable selection	12.62
R^2 of model fitted on all data	0.528
Average R^2 (10-fold cv)	0.485
Standard deviation R^2 (10-fold cv)	0.043
Number of observations	3,450

The *c*-number of the residualised variables was 12.62 before variable selection, which means a mild correlation. The R^2 shows that the model explains about half of the variance, also in 10-fold cross-validation. A model with only random effects reaches an R^2 of 0.478, which again shows that a lot of the variance is explained by the intercepts for Verb and Participant and the random slope for Participant over cCorpusProbs.

The significant coefficients (together with their *t*-values) are shown in Table 5.9. The significant effect for *Variety=US* shows that American participants give higher naturalness scores to prepositional dative constructions than

Table 5.8: Number of values and standard deviation of the random effects in the regression model, fitted on all (combined) judgement data

Random effect	nr of values	stdev.
Participant	115	4.71
Participant/cCorpusProbs		11.21
Verb	9	14.82
Residual		21.81

Table 5.9: Coefficients and their *t*-values for the ratings for the prepositional dative construction by the British, American, and Australian participants, fitted on all data. Insignificant features are in parentheses

Feature	coef	average	stdev	*t*-value	average	stdev
Intercept	48.21	48.24	0.45	9.32	9.27	0.17
AnRec=in	19.06	19.13	0.44	15.05	14.35	0.35
ConTh=in	4.20	4.17	0.40	3.98	3.74	0.37
DefRec=in	14.95	15.00	0.90	7.73	7.36	0.43
DefTh=in	-11.04	-11.06	0.36	-10.48	-9.95	0.33
PrnRec=p	-16.28	-16.38	0.49	-13.56	-13.06	0.48
PrnTh=p	3.72	3.72	0.64	3.02	2.85	0.49
rLenDif	-14.11	-14.13	0.46	-10.03	-9.53	0.30
(Variety=Aus)	1.17	1.17	0.28	0.87	0.84	0.19
Variety=US	3.73	3.72	0.36	2.68	2.60	0.21
rLenDif:Variety=Aus	-3.92	-3.93	0.51	-2.25	-2.14	0.29
(rLenDif:Variety=US)	0.19	0.19	0.83	0.10	0.10	0.44

British participants. Our data revealed no such difference between British and Australian participants (*Variety=Aus*).

With respect to Length difference, we see that Australian participants show a significantly stronger effect than British participants (*rLenDif:Variety=Aus*): as the theme increases in length relative to the recipient, the Australians increasingly favour the double object dative more so than the British participants. This difference could not be found in our American and British data.

5.4.6 Discussion

In the individual models per variety (summarised in Figure 5.2), our data showed that younger British participants tend to be influenced more strongly by the pronominality of the theme than older British participants. In vari-

ous dialects of British English, speakers are relatively flexible with respect to pronominality: They allow reversed double constructions such as *give it him* (Siewierska & Hollmann, 2007; Haddican, 2010). Since our participants come from many regions in the UK (see Figure 5.3), it is likely that many are familiar with this construction.[8] That the effect of the pronominality is getting stronger could indicate that British English is moving away from the marked dialectal construction. We should note that the two-way interaction was only just significant (t=-2.33), which explains why it was not significant as a three-way interaction (with Variety) in the combined model.

Figure 5.3: Google map with the places where the British participants spent most of their youth

As for the effect of age, the individual model for American English revealed that in the US, the prepositional dative construction is most popular with the younger participants. The individual model for Australian English showed a similar effect for younger Australian men. In the combined model, we also dis-

[8] Because of the many different regions present in the British data, including it in the model would hardly differ from including the individual speaker.

covered a difference in the preference for the prepositional dative construction across varieties: American participants gave higher naturalness scores to this variant than British participants, and there was no such difference between British and Australian participants. The intercepts in the three individual models in Figure 5.2 show the same: it is obviously the highest for the US participants. Our corpus study in the previous Section showed no such Variety effect. Recall that to enable comparison between the three varieties, we decided to use the same items for all participants, only slightly adapted to match their nationality (e.g. changing place names). Given the fact that these items were all taken from an American English corpus (Switchboard), our finding for the US participants could be an artefact of the stimuli chosen.

With respect to length difference, our data indicated that Australian participants show a stronger effect than British participants. This difference is not found for American and British participants, making it similar to what was found in the study by Bresnan and Ford (2010) that we extended. In the previous section on corpus data, we found a significant interaction of length difference and Variety in spoken American and British English. This effect has disappeared in the judgement study. The same is true for the significant interaction between Variety and the definiteness of the theme in the corpus study. We will come back to this issue in the next section.

5.5 Discussion and conclusion

Despite the substantial amount of existing research on the dative alternation in English, various questions still remain to be answered. Many comparisons between the dative alternation in different varieties of English are still to be made. Also, the effect of extralinguistic factors on the dative alternation has so far received only little attention in multivariate analyses. In this chapter, we therefore aimed for two research objectives: (1) establishing similarities and differences in the dative alternation in British, American, and Australian English, and (2) establishing similarities and differences in the dative alternation of speakers varying in gender and age. The findings were related to previous research on the dative alternation in different varieties of English, all in a framework of probabilistic linguistics. In that framework, we assume that linguistic factors influence syntactic structure, and investigate their relative contribution to the likelihood of the two constructions in different speaker groups.

The two studies have shown that there are certain patterns in the data sets that are in line with each other and with previous work. These patterns show a *harmonic alignment*: speakers, writers, and participants in experiments tend to

place or prefer phrases with certain characteristics before phrases with other characteristics:

animate usually precedes inanimate
definite usually precedes indefinite
shorter usually precedes longer
pronoun usually precedes nonpronoun

These patterns have also been found for other syntactic alternations (e.g. also in the genitive alternation, Szmrecsányi, 2010), varieties of English (e.g. British, American, Australian, New Zealand, Indian and African-American English), and types of data (speaking versus writing, corpus or experimental). A probabilistic linguistics framework thus seems suitable for modelling effects that certain *linguistic factors* have on syntactic structure.

Besides similarities, the studies have also revealed some distributional differences between varieties and between speakers varying in age and gender, both with respect to the distribution of the two dative constructions and with respect to the relative contribution of the predictive features.

With respect to variety (the first research objective), we used a corpus and a judgement study to find differences and similarities in dative alternation in British, American, and Australian English. American English and Australian English originated from British English at different times in history. Despite the cultural exchange that dominates our modern society, we expected that we would find other distributional differences between British and American English than between British and Australian English. This is indeed what we found.

In our corpus study, we found that in contemporary spontaneous spoken English, American speakers show a stronger tendency towards the double object construction when the theme is indefinite (e.g. *give him a book*) than British speakers. There is a greater preference for the prepositional dative as the recipient length increases relative to the theme, but the effect is less strong for the American speakers. We suggested that the relative importance of these two features may have evolved further in American English than in British English, following the developments found in the diachronic study by Wolk et al. (2012). But seeing that these effects were not replicated in our judgement study, our conclusion is speculative. Our judgement study indicated that the effect of length difference is stronger for Australian participants than for British participants. This is exactly the same as what was found in the study by Bresnan and Ford (2010) that we extended.

We also investigated the differences and similarities in the dative alternation in British, American and Australian English, made by participants varying

in *age and gender* (the second research objective). The British judgements revealed that younger British participants seem to be influenced more strongly by the pronominality of the theme than older British participants. In many dialects in British English, it is common to use double object constructions with pronominal themes, e.g. *give the man it* (Siewierska & Hollmann, 2007; Haddican, 2010). Our study shows that younger participants are more in favour for the prepositional dative variant (*give it to the man*) than the older participants, which means they are moving away from the dialectal construction. The US judgements showed that the prepositional dative construction is most popular with the younger participants, so it seems that this construction is becoming slightly more popular in American English. A similar effect was found in the Australian judgements: especially younger Australian men seem to be more positive towards using the prepositional dative construction. In general, Australian men are more positive towards using the prepositional dative than Australian women, which supports the findings in Bresnan and Ford (2010).

This study has presented novel evidence that there are universal features playing a role in the dative alternation in English, but that there are subtle distributional differences between the roles that they play across varieties, and across speakers varying in gender and age. We have used a corpus study and a judgement study to reach this goal. Seeing that people seem to use language production processes to predict language input during language comprehension (Pickering & Garrod, 2005), we expected to find similar results in the two studies. However, this was not always the case. The results, and comparisons to existing research, could not always be interpreted straightforwardly since the underlying data sets are not fully compatible. For instance, the American corpus data contained spoken telephone dialogues only, while the British corpus data contained all kinds of spontaneous speech except telephone dialogues. Also, despite the high inter-annotator agreement between the annotations of the British and American corpus data, there are always differences (see also Chapter 4, where we also used these two data sets). In the experimental data, our choice for using American-based items for all three varieties may have influenced the results.

There is one more observation we should make. The corpus study revealed that most of the variance in the dative alternation in spontaneous speech can be explained by the verb and the combination of the verb and the theme head. In the judgement study, most of the variance could be explained by the verb of the test item and by the individual participant. In both studies, the predictive features under investigation played a significant, but minor, role. This is often the case in (psycho)linguistic studies (cf. Baayen, 2008). In order to establish the universality of the features, on top of the effect of frequent lexico-syntactic patterns and participant-specific preferences, future research should

be directed at studying the dative alternation and other syntactic alternations in languages other than English.

6

Model interpretation

Edited from: Theijssen, D., ten Bosch, L., Boves, L., Cranen, B., & van Halteren, H. (2012). Choosing alternatives: Using Bayesian Networks and memory-based learning to study the dative alternation. *Corpus Linguistics and Linguistic Theory*. (Accepted for publication.)

Abstract

In existing research on syntactic alternations such as the dative alternation, the linguistic data is often analysed with the help of logistic regression models. In this chapter, we evaluate the use of logistic regression for this type of research, and present two different approaches: Bayesian Networks and Memory-based learning. For the Bayesian Network, we use the higher-level semantic features suggested in the literature, while we limit ourselves to lexical items in the memory-based approach. We evaluate the suitability of the three approaches by applying them to a large data set (>11,000 instances) extracted from the British National Corpus, and comparing their quality in terms of classification accuracy, their interpretability in the context of linguistic research, and their actual classification of individual cases. Our main finding is that the classifications are very similar across the three approaches, also when employing lexical items instead of the higher-level features, because most of the alternation is determined by the verb and the length of the two objects (in Chapter 1: *Snow White* and *the poisonous apple*).

6.1 Introduction

In linguistic research, as introduced in Chapter 1, the goal is to find a model that describes language data accurately, and that tells us something about the roles that certain linguistic features play. The modelling technique commonly used in previous linguistic research, logistic regression, is attractive for several reasons. First of all, it is a multivariate approach: it enables us to investigate the contribution and significance of several features at the same time. Second, contrary to alternative classifiers such as LDA (e.g. Gries, 2003) that make strong assumptions about the statistical distributions of the data, regression models are able to deal with non-numerical data. This is beneficial since nominal (often binary) data is very common in corpus studies on syntax. Third, the models themselves are fairly simple; they provide coefficients that indicate the relative roles that the individual features (values) play. Fourth, multiple regression models make it possible to combine fixed variables (the features) and random variables (random effects). This combination helps to establish the effect of the linguistic variables of interest, while controlling for random variables that are usually not of primary interest, such as the individual speaker.

However, there are also some problems with these regression models. One of the major drawbacks of logistic regression is that it requires certain properties of the data that cannot always be fulfilled. Features should be independent, for instance, but in reality they are are often correlated. For example, it is known that the dative alternation is influenced greatly by the relative lengths of the recipient (e.g. *Snow White* in Chapter 1) and the theme (*the poisonous apple*), but also that humans tend to place pronouns *before* full noun phrases in the clause. These two features, length and pronominality, are correlated because pronominal objects are usually short, i.e. consisting of a pronoun only. Correlated features cause problems with the interpretation of the roles that the individual features play in the model. For example, correlations can cause coefficients to flip sign or loose statistical significance. This means that the effect of pronominality in the model could become insignificant or receive a coefficient that indicates the opposite of the direction expected (on the basis of existing research), because most of its variance is already explained by the length feature. Such correlation issues obviously increase the risk of misinterpreting the effects. There are many mathematical approaches to solve the problems caused by collinearity, for example by centering or residualising variables (for details, see for instance Baayen, 2008). Such approaches mostly involve some form of transformation of the original data into data that has the required characteristics. However, if length difference is for instance residualised on the pronominality of the recipient and the theme (as done in Chapter 5), the feature under investigation is not length difference itself, but

this less straightforward residualised version. Linguists often want to answer research questions about certain features, and transformations tend to hamper the interpretability of the models in terms of the original data. What we thus need are different modelling approaches that do not suffer from these problems.

There is another reason why we want to move beyond regression models. Syntacticians with various backgrounds are now taking more and more interest in the social and cognitive aspects of language. There are, for instance, recent multivariate approaches that combine the results of sociolinguistic studies, researching the effect that factors such as age and gender may have on language, with cognitive linguistic studies (e.g. Geeraerts et al., 2010). Also there are attempts to relate findings from corpus studies to observations in psycholinguistic experiments (e.g. Bresnan & Ford, 2010). Multivariate models such as regression models can successfully be exploited for the purpose of analysing the relative importance of higher-level features (such as those introduced in Chapter 1 and used throughout this thesis) in specific data sets under investigation, but these models cannot elucidate the role of these features in cognitive processes. The first goal of the present chapter is therefore to investigate the explanatory power of the higher-level features in a model that is more likely to be cognitively plausible than regression models. Several approaches have been developed for this purpose, of which connectionist models are perhaps the best-known (e.g. McClelland et al., 2010). Although connectionist models have gained substantial interest in psycholinguistics, they have less traction in formal linguistics, probably because the internal structure of these models is opaque. Recently, Baayen (2011) used Naive Discriminative Learning to model the dative alternation with higher-level features.

In this chapter, we use yet a different approach: *Bayesian Networks* (Pearl, 1988). This approach is fully transparent and does not make assumptions about the statistical distributions of the predictor variables. Bayesian Networks make it possible to integrate possibly uncertain prior knowledge and possibly erroneous empirical evidence of different types and different sources in a consistent probabilistic framework (cf. 6.3.2 for more detail). Integrating partial and noisy sensory input and volatile procedural and semantic memory is what the brain does all the time, especially in the initial stages of the processing where not all information is available yet. Therefore, Bayesian Networks form an attractive analogue for cognitive processes (Chater, Tenenbaum, & Yuille, 2006; Chater, Oaksford, Hahn, & Heit, 2010).

Computational grammar learning models using Bayesian inference have already been shown to be able to learn the dative alternation in a small set of relatively simple, artificial sentences, making use of grammar rewrite rules only, without any higher-level information (Dowman, 2004). The question therefore arises whether the higher-level features are really necessary for explaining

the dative alternation and for generalising from (small) data sets to actual language production. This brings us to our second research goal: to investigate the suitability of a model that can claim cognitive plausibility and that is not provided with higher-level features, but with lexical items. To that end, we adopt a *memory-based learning* approach (Daelemans & van den Bosch, 2005), in which learning is defined as the storage of some sort of representation of experience (cf. section 6.3.3 for more detail). This memory of previous experience is then used to guide actions in new situations. For the dative alternation, this means that humans learn the contextual suitability of the two constructions by storing some representation of the occurrences they produce themselves and hear or read in other people's language use. In the context of current discussions about the existence of an innate, specifically language-related ability, it is interesting to note that memory-based learning has no need to assume an innate language faculty. Language, according to this theory, is learned from input only, making use of the general cognitive abilities that we possess. The underlying idea of this model therefore shows many similarities with exemplar-based models of language processing (Gahl & Yu, 2006), and with for instance data-oriented parsing approaches (e.g. Bod, 2009). When storing all experience with the dative alternation, there is no reason to abstract away from the original input that we hear by defining higher-level features. This makes the role of the higher-level features used in existing research unnecessary and, using Occam's razor, implausible. The only assumption we need to make for studying the dative alternation in the way we do, is that humans have learned the meaning of a number of verbs and the existence of the semantic roles 'recipient' and 'theme'. Memory-based learning does not make assumptions about statistical distributions of the items that are kept in memory.

In order to address our two research goals, we will employ two approaches to model the dative alternation that can be associated with cognitive processes: Bayesian Networks and memory-based learning. For the sake of comparability, we also include the traditional logistic regression models. We evaluate the suitability of the three approaches for studying the dative alternation, on the basis of the following three criteria:

- the quality of the model in terms of classification accuracy
- the interpretability of the model in linguistic research
- the actual classification of individual cases by the model

The remainder of this chapter is structured as follows: In Section 6.2, we describe the data set and the various features used. The modelling techniques are introduced in Section 6.3 and they are evaluated according to the criteria in Section 6.4. The chapter ends with our general discussion and conclusion, provided in Section 6.5.

6.2 Data

6.2.1 Data collection

The data set was extracted from the 100-million-word British National Corpus (BNC Consortium, 2007), following the semi-automatic approach in Chapter 4, as summarised below.

We used a Perl script to extract all sentences with an occurrence of a dative verb, and parsed them with the Functional Dependency Grammar (FDG) parser, version 3.9, developed at Connexor (Tapanainen & Järvinen, 1997). A second Perl script was used to extract all dative constructions from the syntactic parses (152,008 in total), after which we employed two automatic filtering steps. In the first filtering, we used another Perl script to automatically filter out 44,464 candidates to prevent the influence of other types of syntactic variation than those of interest in this research (passive versus active voice, declarative versus interrogative mode, the placement of adverbials, etc.) and to make sure that the features we want to apply later were applicable (e.g. it is not possible to establish the definiteness of the theme if it is a clause, not a noun phrase). In the second filtering, a Perl script was used to remove candidates that were likely to contain parse errors (21,965 in total).

After the filtering, 85,579 dative candidates remained. We next checked over 17,000 candidates manually, removing candidates that contained parse errors and that were not dative constructions. The checked subset contained all candidates from the spoken part of the BNC (>11,000 candidates), and yielded 7,757 confirmed dative constructions. To increase the diversity in the data, we supplemented the spoken material with a random selection from the written material (>6,000 candidates). The resulting data set contains 11,784 instances, of which 7,757 are spoken and 4,027 written, spread over various genres, e.g. public meetings, private conversations, news paper articles and fiction texts.

6.2.2 Medium and length difference

There are two *basic* features that will be used in all models, both in the models that use the higher-level features and the model that uses only lexical items: (1) Medium (spoken or written) and (2) the length difference between the theme and the recipient. Medium is a binary feature that is easy to establish on the basis of the metadata provided in the BNC. Length difference is used as an approximation of syntactic weight, which is known to play a role because of the principle of end weight (Behaghel, 1909).

There are many (often correlated) alternatives for establishing the syntactic

weight (Shih & Grafmiller, 2011), but in this chapter we limit ourselves to a number of variations of the length difference in words. Since the Bayesian Network tool we employ is not able to deal with interval data, we also include several ways of discretisation, leading to a total of six definitions:

- LenDif: theme length in words minus recipient length in words
- lnLenDif: the log of the ratio between these two lengths
- dLenDif5: an intuition-based discretised version of LenDif with 5 levels (i.e. similar lengths, a longer recipient, a longer theme, a *much* longer recipient and a *much* longer theme)
- dLenDif6: a frequency-based discretised version with 6 levels
- dLenDif10: a frequency-based discretised version with 10 levels
- dLenDif78: a frequency-based discretised version with 78 levels

The cut-off points for the intuition-based discretisation were chosen so that if the ratio between the number of words in the two objects was $\geq 1:3$, the longest of the two was considered *longer*, and when the ratio was $\geq 1:4$, it was considered *much* longer. The frequency-based discretisation in resp. 6 and 10 levels was based on the frequency distributions of LenDif in the data set. For the 6-level discretisation, each level had a frequency of at least 1,100 instances, and for the 10-level distretisation, each at least 400 instances. In the 78-level discretisation, each level contained one unique value of LenDif, with the number of instances per level varying from 1 to 3,522.

6.2.3 Verb

It is known that many verbs have a strong preference for one of the two constructions (e.g. Gries & Stefanowitsch, 2004). For this reason, all models take into account the verb used in the dative construction. In the memory-based model, the verb is included in the lexical items, as will become clear in Section 6.2.5. For the regression model and the Bayesian network, the treatment of the verb is explained in Section 6.3. The 46 verbs used in our research are shown in Table 6.1,[1] together with their frequencies in the data set we use.

[1] Many of the dative verbs are not in the parser lexicon as being dative verbs (see Chapter 4), hence the lower number of verbs (46 instead of 76) in Table 6.1. The verb *read* was not found as dative verb in the ICE-GB and Switchboard corpora in Chapter 4, but is found as such in the BNC. We used the same version of the FDG parser (3.9), so apparently the parser allows *read* as a dative verb in certain contexts. Seeing the small number of instances found with *read* (19), however, it seems that these are exceptions.

Table 6.1: Verbs and their frequencies in the data set

give	6974	leave	124	pass	77	throw	48	permit	31
tell	799	lend	120	charge	74	bear	44	deal	30
send	363	cause	113	promise	74	issue	43	advance	23
show	342	write	112	wish	64	award	42	read	19
pay	232	teach	111	grant	62	play	40	vote	13
offer	206	make	98	feed	61	pose	38	forbid	10
do	205	present	92	deliver	59	serve	38		
bring	179	take	91	allocate	54	refuse	36		
sell	158	deny	89	assign	49	accord	35		
owe	152	hand	81	guarantee	48	bid	31		

6.2.4 Higher-level feature extraction

Two of the three modelling techniques employed in this chapter make use of the higher-level features suggested in the literature and introduced in Chapter 1 (the third technique, memory-based learning, uses lexical items only). These higher-level features are often difficult to define and to annotate with high agreement levels between human annotators. We solve this problem by making use of automatic feature extraction, so that that the definitions are clear and the annotations themselves consistent (see Chapter 4). Moreover, Theijssen, van Halteren, et al. (2011a) show that the quality (prediction accuracy) of logistic regression models applied to data annotated with this automatic method is equally good as the models found for data with manual annotations, as long as there are enough data points. Since the data set used in the present chapter is larger (over 11,000 instances) than the largest set (approx. 8,000 instances) included in Theijssen, van Halteren, et al. (2011a), we believe the automatic feature extraction approach to be suitable for the present research.

All instances in the data set were annotated automatically for eight higher-level features, using the feature extraction Perl script in Chapter 4. Compared to the feature set used in Chapter 4, we now leave out the concreteness of the theme (because the automatic feature extraction differed too much from human labellings, see Chapters 3 and 4) and the number of the recipient and the theme (because number did not play a significant role in the manually annotated data in Chapters 2 and 4). The names, definitions and values of the features are the same as introduced in Table 1.1 in Chapter 1, and are repeated in Table 6.2. All higher-level features are binary.

Table 6.2: Higher-level features for which the instances have been annotated automatically

Name	Definition	Values (binary)
AnRec	animacy of recipient	animate, inanimate
DefRec	definiteness of recipient	definite, indefinite
DefTh	definiteness of theme	definite, indefinite
GivRec	discourse givenness of recipient	given, nongiven
GivTh	discourse givenness of theme	given, nongiven
PrsRec	person of recipient	1st/2nd (local), 3rd person (nonlocal)
PrnRec	pronominality of recipient	pronominal, nonpronominal
PrnTh	pronominality of theme	pronominal, nonpronominal

6.2.5 Extracting lexical items

For the memory-based approach, we use two different variants of lexical items as features, being word forms and lemmas. Since the FDG parser provides lemmas in its output, we extracted the lemmas directly from the FDG parses. As already mentioned in Section 6.1, we assume that humans know the meaning of a number of verbs and the semantic roles 'recipient' and 'theme'. As features, we therefore use specific lexical items present in the recipient, the theme and the verb. Consider sentences 1 and 2:

1. I gave a dog biscuit to it.
2. I gave it a dog biscuit.

The word forms extracted from these sentences would be:

- the verb: *V:gave*
- the recipient head: *Rh:it*
- the beginning of the recipient: *Rb:it*
- the theme head: *Th:biscuit*
- the beginning of the theme: *Tb:a*

For the recipient and the theme, we used a Perl script to extract the head from the dependency parses, as well as the first word or lemma (after removing the preposition *to* in prepositional dative cases). The reason for including the beginning of the recipient and theme is that previous research has indicated that definiteness seems to play a role in the dative alternation. Since definiteness is mostly determined by the presence or absence of certain determiners at the beginning of the object, it may well be that it is not the higher-level feature

itself that influences the choice for either syntactic construction, but the presence of certain words or lemmas. We therefore include these lexical items in this model, to see what role they play in the memory-based learning model.

6.3 Modelling techniques

In this section, we elaborate on the three modelling techniques we use: logistic regression, Bayesian Networks and memory-based learning.

6.3.1 Logistic regression

In this approach, we employ the eight higher-level features in Table 6.2 and the length difference, and include them as predictors in a mixed-effect logistic regression model. The Medium (spoken or written) is also included as a predictor, and so are all its interactions with the nine other predictors. The verb of the construction (e.g. *give*) is included as a random factor.

Using the values of the predictors and the verb i we establish a regression function that predicts the natural logarithm (ln) of the odds that the construction C in instance j is a prepositional dative. The regression function is:

$$\ln odds(C_{ij} = 1) = \alpha + \sum_{k} (\beta_k V_{jk}) + e_{ij} + r_i \ . \tag{6.1}$$

The α is the intercept of the function. The terms $\beta_k V_{jk}$ contain the weights β and values V_j of the predictors k. The random effect r_i established for the verbs (i) is normally distributed with mean zero ($r_i \sim N(0, \sigma_r^2)$), independent of the normally distributed error term e_{ij} ($e_{ij} \sim N(0, \sigma_e^2)$). The optimal values for the function parameters α, β_k, r_i and e_{ij} are found with the help of Maximum Likelihood Estimation.[2]

The variable selection method is as follows: We start out with a model including all predictors and two-way interactions with Medium, and remove all insignificant *interactions* in one single step. This step is carried out on the full set of 11,784 instances available, after which only significant predictors remain. We perform this six times, each with a different representation of the length difference. The discrete representations of length difference are interpreted as binary features: one binary feature for each discrete level (except one that is included in the intercept). The representation with 78 levels (dLenDif78), and hence 77 binary features, runs into sparseness problems,[3] but the other five all score *model fit* accuracies that do not differ significantly from each other,

[2] We use the function `lmer()` (Bates, 2005) in `R` (R Development Core Team, 2008).

[3] The function `lmer()` cannot cope with numerous missing feature value combinations, which is the case with dLenDif78: for 30 of the 78 values, there are <= 3 data instances.

training and testing on the same 11,784 instances: 92.2% to 92.5%.[4] Since the log of the ratio between the two lengths (lnLenDif) scores the highest model fit (92.5%), and is one of the two most parsimonious definitions with respect to number of regression coefficients (only 1, because it is numerical), it is selected for further analysis.

6.3.2 Bayesian Network

The higher-level features that are implicated in selecting a dative construction can be considered as just as many modules in a very complex system that generates sentences. To avoid making things overly complex, we assume that the structure of that system can be represented in the form of an acyclic directed graph, which means that (parent) module M_x can affect the operation of (daughter) module M_y, but not the other way round. Obviously, the fact that we know the direction of the dependencies implies that we claim to have prior knowledge about the structure of the process that we are investigating. It also means that we can draw a picture of the structure of the system in which the modules are represented by nodes, and the connections between the modules are represented by single-headed arrows (cf. Figure 6.1). An arrow from parent node N_x to daughter node N_y implies that the latter is conditionally dependent on the former: The value of N_x influences the operation of N_y. The beauty of the theory of Bayesian Networks (Pearl, 1988) is twofold. First, this theory allows us to learn the strength of the connections between modules from data, which is tantamount to integrating knowledge (represented by the nodes and connections in the network) and empirical observations. Second, the theory allows for efficient computation, because it follows from the structure of the network which modules are conditionally independent. In technical terms: it allows us to factor the joint probability $p(N_1, \cdots, N_m)$ of observing specific values for all m modules (represented as nodes) in the network at a given time in such a manner that only conditional dependencies (represented by arrows in the network) need to be taken into account.

Compared to logistic regression, Bayesian Networks have several advantages. One, which is not relevant in the case of dative alternation, is that the output node (the black node labelled 'Cons' in Figure 6.1) can take an arbitrary number of values. Second, the structure of the Bayesian Network represents a decomposition of the original modelling problem into smaller subproblems. Third, the way in which the features interact is easier to visualise. There are several public-domain software packages that allow one to easily create and manipulate networks and to visualise the strength of the connections that

[4] The accuracy reached when training and testing on the same data is sometimes also referred to as *model fit*, *empirical fit* or *performance ceiling*.

were learned from training data (e.g., in the form of the thickness of the arrows, cf. Figure 6.2). Unsurprisingly, these advantages come at a cost: There is no proven method for learning the *structure (topology)* of a network from training data. Incomplete prior knowledge may cause mistakes in drawing the connection scheme and thus result in misleading accounts of the structure of the process that generates the output observations. For this reason we will explain the decisions that were made in creating the network in Figure 6.1 in substantial detail.

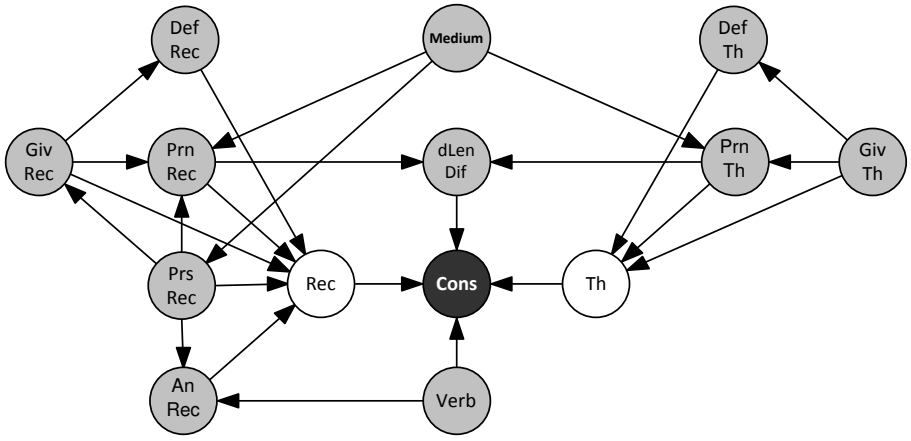

Figure 6.1: Theoretically motivated Bayesian Network. The grey nodes are observable input nodes, the white ones are hidden nodes, the black node is the (observable) output node.

The names of the nodes are the features in Table 6.2, supplemented with a node for Length difference (dLenDif) and a node for Verb, which is treated as a discrete variable with 46 (nominal) values. Since the syntactic construction is queried using Bayesian inference on the evidence set for the other nodes, it is indicated with the black *output* node labelled *Cons*. The network contains two hidden nodes in white: *Rec* (for *Recipient*) and *Th* (for *Theme*). These are nodes that have no values in our training or test data, but are nodes that allow the network to combine and 'summarise' the information about the theme and the recipient in a number of states. This is an elegant way to combine the various, possibly correlated, characteristics of the theme and the recipient (the grey *input* nodes), and see the relative influence they have on the recipient and the theme (i.e. the hidden nodes) separately.

Each of the arrows in the network is motivated below (sorted alphabetically by node name). It goes without saying that many arrows could be added

and removed, either randomly or on the basis of other linguistic intuitions or theories. However, our goal is not to perform data exploration and find the one best model, but to apply our hypotheses about the dependencies between the features and the syntactic construction used, preferring a network structure that is interpretable and transparent.

Animacy (AnRec)

The animacy of the recipient has no direct influence on the other nodes, as far as we know. Therefore, there is only an arrow from AnRec to the hidden node Rec.

Definiteness (DefRec, DefTh)

As far as we are aware, the definiteness has no direct influence on the other feature nodes, so it only has an arrow towards the hidden nodes Rec and Th.

Length difference (dLenDif)

The length difference is known to have a strong influence on the construction used (e.g. Chapter 4, Bresnan et al., 2007), because of the principle of *end weight* (Behaghel, 1909). Therefore, we added a direct arrow from dLenDif to Cons.

Discourse givenness (GivRec, GivTh)

When an object has been mentioned previously in the discourse, we expect that the speaker or writer is more likely to use a pronoun (e.g., referring to a previously mentioned book with *it*), hence the arrows to PrnTh and PrnRec. When the object represents new information, we assume it is more likely to be realised as an indefinite noun phrase (e.g., *a book* is usually a book that has not been mentioned before). Besides arrows to these two features, we also include arrows to the hidden nodes Rec and Th.

Hidden nodes (Rec, Th)

The two hidden nodes receive and collect information from the various feature nodes, and both provide information to Cons.

Medium

Biber (1988) has already shown that spontaneous, usually spoken language contains significantly more pronouns and mentions of first and second persons

(*me*, *you*) than more formal and written language. We therefore added arrows from Medium to PrnRec, PrnTh and PrsRec. The bias towards the double object construction is usually stronger in spoken data (86.6% of the spoken instances in our data set were double object) than in written data (64.3% of the written instances in our data). However, we inspected various existing dative data sets (e.g. those in Chapters 2 and 4) and discovered that these differences can all be explained by the relative frequencies of the values for the three features PrnRec, PrnTh and PrsRec. For this reason, there is no direct arrow from Medium to Cons.

Pronominality (PrnRec, PrnTh)

As mentioned in Section 6.1, the length difference between the two objects is greatly influenced by the pronominality of these objects. The reason is that pronominal objects are often very short because they usually consist of a pronoun only. For this reason, the network includes arrows from PrnRec and PrnTh to dLenDif.

Person (PrsRec)

When the recipient is in first or second person (local), it is almost always an animate, pronominal, discourse given recipient (*me*, *us*, *you*). Because of this direct influence, the network includes arrows from PrsRec to AnRec, GivRec and PrnRec. There is also an arrow to the hidden node Rec.

Verb

As mentioned previously, many verbs have a strong preference for one of the two constructions (Gries & Stefanowitsch, 2004). Therefore, there is a direct arrow from Verb to Cons. Also, some verbs may influence the likelihood of the animacy of the recipient. For instance, in various existing dative data sets (e.g. those in Chapter 2 and 4), we saw that the verb *show* is more likely to occur with an animate recipient, since one usually shows something to people, not to things. This explains our choice to include an arrow from Verb to AnRec.

The network was designed in the Windows user interface GeNIe, a modeling environment for graphical decision-theoretic models developed by the Decision Systems Laboratory of the University of Pittsburgh.[5] The parameter learning on the training data and the inference on the test data was performed in GeNIe's underlying reasoning engine SMILE (Structural Modeling, Inference, and

[5] See `http://dsl.sis.pitt.edu`.

Learning Engine). SMILE is a library of C++ classes implementing graphical decision-theoretic methods such as Bayesian networks and influence diagrams.

Since the goal of the present chapter is to present an overall evaluation of the suitability of Bayesian Networks for modelling the dative alternation, we decided not to perform any tuning of the tool, employing GeNIE/SMILE's default settings instead. By default, the parameter learning is done with Expectation Maximisation with randomised initial parameter settings. For each test case, the evidence of the nodes was set to the feature values in question, after which the beliefs in the network were updated through the default inference approach (the clustering algorithm). The probability assigned to the node *Cons* was then used to classify the case, choosing the class with the highest probability in the histogram provided for the two possible outcomes.

With respect to LenDif, GeNIe/SMILE was able to deal with the discretised versions only, hence the label *dLenDif* (with a *d* for discretised) in Figure 6.1. For the hidden node *Th*, we tested all seven possible numbers of values, given the binary input from the three parent nodes: 2, 3, 4, 5, 6, 7 and 8. The same numbers were tested for *Rec*, supplemented by 16, 24 and 32 because of the higher number of parent nodes (5) and hence the high number of possible input combinations ($2^5 = 32$). To explore the effect of different cardinalities of the nodes Th, Rec and dLenDif, we thus tried 7 (Th) × 10 (Rec) × 4 (dLenDif) = 280 combinations. Note that all models are the same in their network structure; they only differ with respect to the number of values possible for some nodes. To find the optimal settings, we learned and predicted the same data set with all 11,784 instances.[6] There were 159 combinations that yielded prediction accuracies which did not differ significantly from the highest accuracy (95.1%), i.e. they were all within the 95% confidence intervals according to a binomial distribution. Two of these combinations represented the most parsimonious representation (requiring only 12 values in total): 5 for dLenDif, 4 for Rec and 3 for Th, and 6 for dLenDif, and 3 for both Rec and Th. We only present the results for the former, since it yielded the highest accuracy (94.5%, compared to 94.3% for the latter option).

6.3.3 Memory-based learning

Logistic Regression and Bayesian Networks have in common that they use the training data to learn generative models that, given the values of a set of parameters of a new observation, can predict the class to which that observation belongs. Learning the models requires substantial effort and expertise, more often than not expertise at a level that cannot reasonably be expected from naive language users. For example, in this chapter we do not include the feature

[6] Again, we thus established the *model fit*.

'Concreteness of the theme', which does appear in Bresnan's model (Bresnan et al., 2007), because of the problems we experienced in annotating that feature (see Chapter 3). Generative models also run into trouble if some feature values occur rarely in the training data (cf. section 6.3.1.) Memory-based learning as defined by Daelemans and van den Bosch (2005) is a machine-learning method that is designed to avoid problems with labelling data on an abstract level, as well as with sparse observations. Memory-based learning does exactly what its name says: Training examples are stored in the form in which they are observed in text or speech. The only mandatory annotation is the label of the class of which the examples are a member. All training examples are characterised by a number of simple, theory-neutral features, such as the identity of words in a phrase, the identity of the left and right neighbour of a word, the number of syllables of a word, etc. When a new observation comes in to be classified, the examples stored in the memory are searched for items that are most similar (in terms of the features) to the new observation. Learning now consists of finding the similarity measure that minimises the classification error for the training data. Because it does not rely on any kind of generative model, memory-based learning can deal with low-frequency events, even if these represent sub-regularities.

For memory-based learning, we included the two types of lexical items described previously, together with the Medium and one of the six versions of length difference (each version tested in a separate model). The implementation we employed is the nearest neighbour (kNN) classifier in TiMBL (Daelemans, Zavrel, van der Sloot, & van den Bosch, 2010). TiMBL stores classified (training) data, and the items in the test set are assigned the class of the nearest neighbour in the stored data. We used the leave-one-out setting, which is a procedure of iteratively training on all-but-one instances, and testing on the one remaining instance.

TiMBL can be tuned by setting a number of hyperparameters, including the distance metric used for each feature (m), the feature-weighting method (w), the number of nearest neighbours used for extrapolation (k) and the type of class voting weights that are used for extrapolation from the nearest neighbor set (d). For each of the twelve lexical item/LenDif variants, we separately tuned these hyperparameters with the help of the wrapper Paramsearch (van den Bosch, 2004). Paramsearch finds the best settings by cleverly trying out parameter combinations on subsets of the data. We provided Paramsearch with all data instances and saved the settings that were chosen as 'optimal'. These settings were next used in the leave-one-out setting mentioned above: m = Jeffrey divergence, w = Gain Ratio, $k = 9$ and d = normal majority voting (i.e. all neighbours have equal weight).

All combinations of the type of lexical item (lemma or word) and the defi-

nition of length difference yielded an accuracy between 92.4% and 93.1% when training and testing in leave-one-out mode. Since the lemma-based features are more parsimonious (5,563 different lexical items) than the word-based features (6,358 different lexical items), we focus on the lemma-based models. From these, we have selected the model that yielded the highest accuracy (93.1%) for further analysis, which was the model using the discretised version of length difference with 10 levels (dLenDif10).

6.4 Evaluating the approaches

6.4.1 Quality of the model in terms of classification accuracy

We evaluate and compare the predictions made by the various models by using the models as classifiers and establishing the percentage of correctly classified instances (the accuracy). We did this in two ways: (1) training and testing on all instances (leave-one-out for Memory-based learning), yielding the model fit, and (2) training and testing in 10-fold cross-validation, using the same division in 10 folds across the approaches. In the 10-fold cross-validation, we re-used the output of the variable selection and hyperparameter tuning applied to all data instances (as described in the previous Section).[7] The model fit accuracies and the average 10-fold accuracies reached can be found in Table 6.3.

Table 6.3: Accuracies and their confidence intervals (for model fit) or two times the standard deviations (for 10-fold cross-validation), found for the two baselines and the three modelling approaches

Approach	Features	Model fit	10-Fold cv
Class-majority baseline	none	79.0% (±0.7%)	79.0% (±2.1%)
Verb/LenDif baseline	basic	89.6% (±0.6%)	89.3% (±2.4%)
Logistic regression	basic+higher-level	93.5% (±0.4%)	93.2% (±1.2%)
Bayesian Network	basic+higher-level	94.5% (±0.4%)	93.2% (±1.3%)
Memory-based learning	basic+lexical	93.1% (±0.5%)	92.5% (±1.5%)

For the model fit (leave-one-out for memory-based learning), the three models perform much better than the class-majority baseline of 79.0% (always selecting the double object construction). They are only slightly, but significantly, more

[7] Strictly speaking, this is not a fair train-dev-test split, since we tune on the complete data set (including test data). But since our qualitative evaluation will be based on the models built on all instances, we wanted the variables and parameters of the 10 models in the cross-validation to match those of these models. We believe this decision is defendable because all three approaches have the same benefit.

accurate than the Verb/LenDif baseline, using the verb and length difference only (89.6%).[8] As mentioned previously, many verbs have a strong preference for one of the two constructions (e.g. Gries & Stefanowitsch, 2004). Also, the length difference, which could be interpreted as an approximation of the principle of end weight (Behaghel, 1909), is known to have great influence.

When training and testing on all items (the model fit), the best results are reached with the Bayesian Network using the higher-level features (94.5%). In 10-fold cross-validation, the three approaches do not differ significantly, yielding accuracies of 92.5% or higher. The standard deviations for the three approaches are remarkably smaller than those found for the two baselines, which means that the addition of higher-level features or lexical items has led to more stable models. It is interesting to see that a memory-based model, which uses only the basic features and lexical items, is so accurate at predicting the construction used. This is a reason to call into question the importance of higher-level features in language processing. Also, it adds to the questioning of the need for an innate, specifically language-related ability, since memory-based learning explicitly assumes that language is learned from input only, making use of the general cognitive abilities that we possess.

As mentioned in Section 6.1, the goal in linguistic research is not to find the best performing model, but to find an approach that is sufficiently accurate to constitute a plausible explanation of the underlying cognitive processes and that, at the same time, is able to teach us something about linguistics. The models that we investigate all show a high accuracy; therefore, we keep all three in a more qualitative evaluation.

6.4.2 Interpretability of the model in linguistic research

In this section, we will evaluate the interpretability of the models in linguistic research, treating them each in a separate subsection.

Logistic regression

The coefficients found for the fixed factors in the logistic regression model are presented in Table 6.4. What we can learn from the model is that all predictors are significant except Medium, which is kept in because of its significant interaction with DefTh. The fact that so many predictors are significant is not surprising given the large number of data instances. The coefficients in the model can be interpreted because they directly influence the log of the odds

[8] This score was reached with a logistic regression model with verb included as a random effect and length difference (dLenDif5) as the only fixed factor. The type of length difference had no influence on the accuracy reached. Memory-based learning and Bayesian Networks also scored accuracies above 89.0% when provided with only the verb and a form of length difference.

that the construction used is the prepositional dative. So, if the recipient is inanimate, the odds increase with 1.03. On the other hand, if the recipient is pronominal, the odds decrease with 1.29, thereby increasing the odds that the construction used is the double object.

Table 6.4: Coefficients and their properties in the logistic regression model

Feature	Coefficient	Std error	z value	$Pr(>\|z\|)$	
(Intercept)	1.14	0.39	2.93	0.003	**
AnRec=in	1.03	0.11	9.37	0.000	***
DefRec=in	0.92	0.14	6.79	0.000	***
DefTh=in	-1.23	0.16	- 7.67	0.000	***
GivRec=non	0.86	0.14	6.10	0.000	***
GivTh=non	-1.44	0.15	- 9.37	0.000	***
LenDif	-2.30	0.08	-27.16	0.000	***
PrnRec=p	-1.29	0.15	- 8.67	0.000	***
PrnTh=p	1.32	0.12	10.78	0.000	***
PrsRec=non	0.33	0.12	2.68	0.007	**
Medium=w	-0.05	0.14	- 0.33	0.741	
DefTh=in, Medium=w	0.55	0.17	3.12	0.002	**

Our regression model confirms that animate objects are usually mentioned before inanimate objects, definite before indefinite, discourse given before discourse new, shorter before longer, pronominal before nonpronominal and local (1st/2nd person) before nonlocal (3rd person). As mentioned in Section 6.1, the fact that regression models are fairly straightforward is one of the reasons that they have become so popular among syntacticians studying alternations.

It is unclear, however, how the model has dealt with the correlations between the features. The collinearity in the data can be measured with the help of the condition number (c-number). For the features in our data, the c-number[9] is 14.20, which indicates that there is medium collinearity. In models fitted to smaller data sets, effects of collinearity can become apparent because not all features reach significance. For a large data set such as ours, this is not the case: all features (except Medium) are highly significant. Collinearity can also cause coefficients to flip sign: if two predictors are (strongly) correlated, the predictor with the highest correlation with the criterion will leave only a residual to explain by the predictor with the weaker correlation. The correlation with the residual may have the opposite sign. Seeing that the patterns found are consistent with those found in the vast body of research (including studies using experimental data, and studies investigating the features one

[9]We used `collin.fnc()` in the `languageR` package in `R`.

at a time), it seems there is no clear influence of collinearity. Still, comparing the actual values of the coefficients, and thereby the relative influence of the feature on the construction used, is not advisable. Another motivation for refraining from a comparison of the coefficient sizes is the fact that most of the statistical variance is explained by the random effect verb and the feature length difference, reaching a model fit accuracy of 89.6%. This means that the coefficients for the other features have only a minor influence on the eventual classification.

Bayesian Network

The Bayesian Network that we used was already presented in Figure 6.1. In the user interface GeNIe, it is possible to calculate the strength of influence per arrow, and represent this visually in the network. By default, this strength is considered equivalent to the extra information obtained by knowing the value of the parent, compared to the situation where this information is not available. Since in our case each node is characterized by a discrete probability density function (pdf) specifying the probability for, say, N different values, the strength can be represented as the Euclidean distance $E(Node, Parent)$ between the conditional probability of a node given the parent node and the a-priori probability of the node (Koiter, 2006):

$$E(node, parent) = \frac{\sqrt{\sum_{n=1}^{N} \left(P_n(node|parent) - P_n(node)\right)^2}}{\sqrt{2}} \ . \qquad (6.2)$$

where $P_n(.)$ represents the n^{th} component of the discrete pdf $P(.)$. Since the minimum value of the sum is equal to 0 and the maximum to 2, the division by $\sqrt{2}$ ensures that the resulting distance is between 0 and 1. The strength of influence thus represents a kind of 'local information gain' yielded by the evidence provided by the parent.

The strengths are shown in Figure 6.2 by the thickness of the arrows. The figure shows that many of the correlations between the features show thick arrows, indicating they are strongly determined by the value of their parent nodes. This is exactly what we expected. Also, we see that the influence of the features on the two hidden nodes Rec and Th has a very similar strength across these features. It therefore seems that the features are similar in their informativeness for the hidden node, and that the hidden node indeed nicely summarises the information of various correlated features. There are only minor differences in the thickness of the arrows: for both Rec and Th, for instance, the node for givenness (GivRec or GivTh) is one of the more influential.

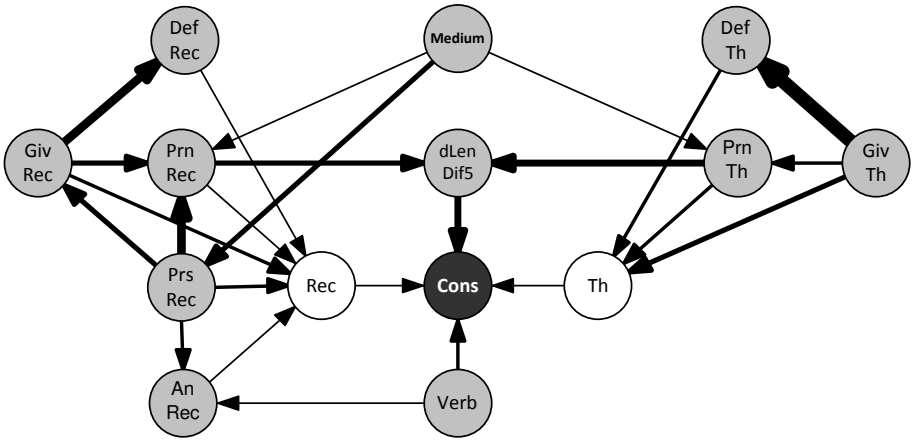

Figure 6.2: Strength of influence in the Bayesian Network (parameters learned on all data). The thickness of the arrows represents the 'local information gain' of knowing the value of the parent node (for more details, see the text).

When we look at the output node Cons, we see that the characteristics of the recipient and the theme have a weaker influence on Cons than the verb and the length difference. This is as we expected, especially after seeing the high score for the Verb/LenDif baseline in Table 6.3.

We should note that although the thickness of the arrows nicely visualises which nodes are strongly determined by their parent nodes, the thickness only represents the strength of influence at that (local) place in the network. Consequently, the strengths in Figure 6.2 do not indicate which arrows are most relevant in the classification task (predicting which Cons was used). For this reason, we established the model fit accuracy, i.e. training and testing on all 11,784 instances, of networks in which we removed one of the arrows. This procedure revealed that the accuracy only dropped significantly when removing one of the four arrows connected directly to Cons. Removing any of the other arrows, even the very thick ones such as that from GivTh to DefTh, did not yield a model fit accuracy that differed significantly from the original network with that arrow. This confirms the general observation that many of the features (here represented as nodes) overlap in their explanatory power: for a large part, they provide the same information. The model fit accuracy decreased most when removing the arrow from dLenDif to Cons (namely to 91.1%), closely followed by the arrow from Verb to Cons (91.5%) and from Th to Cons (91.6%). Removing the arrow from Rec to Cons led to an accuracy of 92.5%. The fact that Verb and dLenDif are very informative is not surprising,

seeing our findings in the previous sections.

Memory-based learning

Compared to logistic regression and Bayesian Networks, the memory-based learning model does not allow an easy interpretation at a more general and abstract, linguistically meaningful, level. The only thing we can deduce from the TiMBL output is the Gain Ratio and Information Gain of the *individual* basic and lexical features, as provided in Table 6.5. The Information Gain measures the difference in uncertainty (i.e. the entropy) between the situation where the feature value is known, and the situation where only the a-priori probability of the class (the dative construction) is known. It is thus very similar to the influence strengths in the Bayesian Network. The Gain Ratio is based on the Information Gain, but normalises it for features with different numbers of values (by dividing the Information Gain by the entropy of the feature values). Only the Gain Ratios are actually used in the model, i.e. as weights in the feature-weighting metric selected in the hyperparameter tuning. The features in Table 6.5 are therefore sorted according to their Gain Ratios.

Table 6.5: Individual features, their number of values, Gain Ratios and Information Gain (provided) in the memory-based model (trained on all data)

Feature	Nr of values	Gain Ratio	Information Gain
dLenDif10	10	0.097	0.275
Rh	1,464	0.067	0.350
Rb	888	0.063	0.275
V	46	0.050	0.149
Medium	2	0.050	0.047
Tb	1,032	0.048	0.257
Th	2,133	0.040	0.367

When we look at the Gain Ratios, we see that the length difference receives the highest feature weight. The verb (*V*) ends only in the middle of Table 6.5 in the ranking for Gain Ratio, and only Medium has a lower Information Gain. This is surprising since we know from previous research (e.g. Gries & Stefanowitsch, 2004) and from the Verb/LenDif baselines that the verb is very informative.

The Gain Ratios reveal that especially the characteristics of the recipient weigh heavily in the classification; they are ranked above all other lexical features. So, despite the many possible values for the two features for the recipient (1,464 and 888), knowing the beginning and/or the head lemma is informative. The reason that both recipient features have a high Gain Ratio

is probably that for 9,519 instances (80.8%), the recipient consists of one word only, which means that the features *Rh* (head of the recipient) and *Rb* (beginning of the recipient) have the same value: this one word that is the recipient. Of these, 8,465 instances (71.8% of all data) have a recipient that is the personal pronoun *you*, *me*, *them*, *him*, *us*, *her* or *it*. The beginning and the head lemma of the recipient therefore give information about the pronominality (and probably also the short length) of the recipient. The high Gain Ratios thus seem to confirm the finding in previous research that pronominality plays a role in the dative alternation.

The Information Gain values for the two features for the theme are very close to the ones found for the recipient. However, when looking at the Gain Ratio, which takes into account the many values the features can take, we see they are not so informative compared to the other features.

In our description of the lexical items used, we explained that we wanted to include the beginning of the recipient and the theme in order to test whether the relevance of definiteness found in previous research can be explained with the help of lexical items. Table 6.5 shows that both features representing the beginning of the objects (Rb and Tb) are quite informative with respect to Information Gain, but only Rb also receives a relatively high feature weight (a Gain Ratio of 0.063, ranked third, compared to a Gain Ratio of 0.048 for Tb, ranked sixth). Based on our observations above, we believe that for the *recipient*, the higher Gain Ratio is most likely caused by the pronominality (and possibly also the length) of the recipient, and not so much by the definiteness. The two most frequent beginning lemmas of the *theme* are the two English articles *a* (3,219 instances, 27.3% of the data) and *the* (1,593 instances, 13.5%). However, since the model output only provides Information Gain and Gain Ratio scores for complete features, and not for the individual feature values that provide information about definiteness, it is not possible to draw any conclusions about the role of definiteness in this memory-based model.

Despite the fact that the memory-based model is difficult to interpret in the sense of understanding which lexical items are most relevant for the choice between the two dative constructions, the model is still useful in the context of linguistic research. Many researchers believe that humans learn language by storing examples, without abstraction in the way it was suggested in traditional linguistic research. Our memory-based model helps to increase the plausibility of this theory.

6.4.3 Classification of individual cases by the model

Besides evaluating the quality of the models in terms of classification accuracy, and their interpretability in linguistic research, it is interesting to compare the

actual classifications made by the models, because they reflect the differences between the models. We do this in two ways: (1) by comparing the classes assigned to the cases, and (2) by comparing the confidence scores provided with these classes.

Comparing the classes

The four panels in Table 6.6 show four different confusion matrices: one for the 10,837 instances (92.0%) that received the same class from all three approaches, one for the 241 instances (2.0%) for which the class found with Logistic regression differed from the other two, one for the 143 instances (1.2%) for which the Bayesian Network differed, and one for the 562 instances (4.8%) for which Memory-based learning differed. Since the classification problem is binary and we tested only three classification approaches, all data points are covered in the confusion matrices.

Table 6.6: Confusion matrices of the 11,784 double object (DO) and prepositional dative (PD) instances for which the construction was predicted (Pred=DO or Pred=PD)

a: 10,838 (92.0%) instances classified the same by the three approaches

	Pred=DO		Pred=PD	
DO	8,715	80.4%	116	1.1%
PD	216	2.0%	1,791	16.5%

b: 562 (4.8%) instances classified differently by Memory-based learning

	Pred=DO		Pred=PD	
DO	96	17.1%	162	28.8%
PD	200	35.6%	104	18.5%

c: 241 (2.0%) instances classified differently by Logistic regression

	Pred=DO		Pred=PD	
DO	7	2.9%	146	60.6%
PD	48	19.9%	40	16.6%

d: 143 (1.2%) instances classified differently by the Bayesian Network

	Pred=DO		Pred=PD	
DO	36	25.2%	28	19.6%
PD	41	28.7%	38	26.6%

The confusion matrices show that most instances (10,506) receive the same, correct, class in the three approaches: 8,715 double object (DO) cases and 1,791 prepositional dative (PD) cases. So, despite the different modelling techniques of the three approaches, and the different types of features used (lexical and higher-level), the vast majority of the instances is classified correctly in all three approaches. This is not surprising since 89.6% (see Table 6.3) was classified correctly with Verb and LenDif only, which were both present in the three approaches as well. In fact, of the 10,506 instances that were correctly classified by the three approaches, 94.9% (9,971 instances) was also classified

correctly by the Verb/LenDif baseline.

Of the 116 double object (DO) instances that were classified as prepositional dative (PD) constructions by all three approaches, 46 were instances where both the theme and the recipient consisted of a pronoun only (see examples 3, 4 and 5). In total, our data set contained 95 of such double object instances, of which only 12 were correctly predicted by all three approaches. The reason probably is that in examples 3 and 5, the alternative (e.g. *give it to you*/*him*) is also very common, making it hard to learn when humans use which.

3. If we give you that we can *give you it* in a certain way, but it is not necessarily meaningful. (BNC: FUL n1285)
4. but you can always say no to any pack you don't want, you're never under any obligation to buy and we'll stop *sending you them* whenever you ask (BNC: HKD n20)
5. Well they won't *give him it* straight away, they'll see to you first. (BNC: KCX n1835)

Memory-based learning differs most from the other two approaches (562 instances), and most of these differences lead to misclassification (362 instances). The misclassifications comprise a relatively large proportion of instances containing recipients that are non-pronominal (70.2%), in third person (85.4%), non-given (53.6%) and/or inanimate (35.4%), compared to the rest of the data (24.5%, 51.6%, 20.0% and 14.6%, respectively). Objects in these semantic categories can be instantiated by a much larger number of different words than objects that are pronominal (usually simply one of the pronouns), in first or second person (*me*, *us* and *you*), given (from the limited set of previously mentioned entities) or animate (a person or animal). Since the memory-based learning model makes no use of the higher-level features, but only of dLenDif, Medium and lexical features, it is not very surprising that it performs best at the instances with objects instantiated by more frequent words.

It remains unclear whether the memory-based model fails to classify the more unique instances correctly because of its inability to abstract away from the data (while humans may in fact be doing so), or because its exposure to language data is too small (especially compared to the amount of language to which humans are exposed). Moore (2003) estimated that infants hear approximately 6 million words of speech a year, and adults approximately 14 million. The data we presented to the model was extracted from a corpus of 100 million words. Since we only checked around 20% of the dative candidates found by the parser, the data set could be taken as representative for approximately 20 million words of the corpus. These are words in speech and writing, while the

estimated quoted from Moore (2003) was speech only. We can thus safely say that humans, over the years, hear many more dative sentences than the 11,784 we used in the memory-based learning approach.

Logistic regression differs from the other two approaches in 241 cases, most of which are instances where a DO construction is wrongly classified as a PD construction (146 instances). Over 70.5% of these misclassified DO cases were taken from written data, while the percentage of written instances is only 33.7% in the data used in this study. It thus seems that the Logistic regression model is especially tailored towards spoken data (the larger part of the data).

The classification by the Bayesian Network differs least from the other two approaches. The 143 differing cases are spread relatively uniformly in the confusion matrix, showing no clear pattern as to where and why the classification differs.

Comparing the confidence scores

The classifiers not only assign a class label to each case, but also a measure of confidence. In order to compare these measures, we transformed them so that all three represent the likelihood that the construction used was prepositional dative. For Bayesian Networks, we took the probability for the prepositional dative from the histogram provided by GeNIe. For regression, we transformed the log of the odds that the construction was prepositional dative into probabilities. For the memory-based learning models, we used the normalised distributions given in the model output,[10] being values between 0 and 1. The higher this value, the higher the proportion of prepositional datives in the set of nearest neighbours. The three transformed confidence scores will from now on be referred to as *PD-likelihood scores*.

Table 6.7 presents the pairwise Pearson correlations for the PD-likelihood scores assigned by the three classification models. The three correlations are all ≥ 0.88 (indicating high correlation) and highly significant ($p < 0.001$). We should note that these high correlations are mostly the result of the fact that the larger part of the data has PD-likelihood scores close to 0 and to 1. The correlations with Memory-based learning are lowest, which shows that the likelihood scores differed most in this approach. There are two possible explanations for this finding: (1) the type of input features used (lexical vs. higher-level) has influenced the PD-likelihood scores, and/or (2) the distribution of the PD-likelihood scores in the memory-based model is different because the scores are proportions, not probabilities. The proportions differ from probabilities especially because they contain many 0's and 1's, while the probabilities only approximate 0 and 1.

[10] We ran Timbl with `+v db -G0` to obtain these normalised distributions.

Table 6.7: Pearson correlation between the PD-likelihood scores assigned by the various approaches (for all $p < 0.001$)

	Logistic regression	Bayesian Network	Memory-based learning
Logistic regression	1.00	0.95	0.88
Bayesian Network		1.00	0.89
Memory-based learning			1.00

It is to be expected that the PD-likelihood scores assigned to cases that were classified correctly are more at the extremes of the likelihood range (close to 0 and 1), while the scores for cases classified incorrectly are more in the middle (around 0.5). To test if this is true for the three models, we established the average likelihood scores for correctly and incorrectly classified DO and PD constructions. These are presented in Figure 6.3.

The boxplots in Figure 6.3 show the expected pattern, and are quite similar for the three models. For the correctly classified double object constructions (the bulk of the data), the mean of the PD-likelihood scores is very low, and the quartile boxes very small. This shows that on average, the three models are very certain that the instance is a double object construction. The quartile boxes for the correctly classified prepositional datives, are much broader. This suggests that the confidence of the classifiers is related to the number of positive training examples that are available. Put differently, the confidence for the majority class is higher because the a-priori probability of correct classification is already much higher. At the right hand side of the figure, the two groups of cases that were misclassified receive scores that are approximately equally close to the extremes (0 and 1) as to the middle (0.5). So, despite the fact that the models classified these instances incorrectly, they are fairly certain about the classification, though not as certain as for the correctly classified cases. The likelihood scores are especially extreme for the Memory-based learning model; apparently, in most cases a large proportion of the nearest neighbours represents one of the two dative constructions, which is then selected as the class for the test item.

6.5 General discussion and conclusion

In this chapter, we have compared three different approaches to modelling the dative alternation. The first approach was one that is commonly used in linguistics: logistic regression models combining various higher-level features. The second approach used the same features, but a modellling technique that

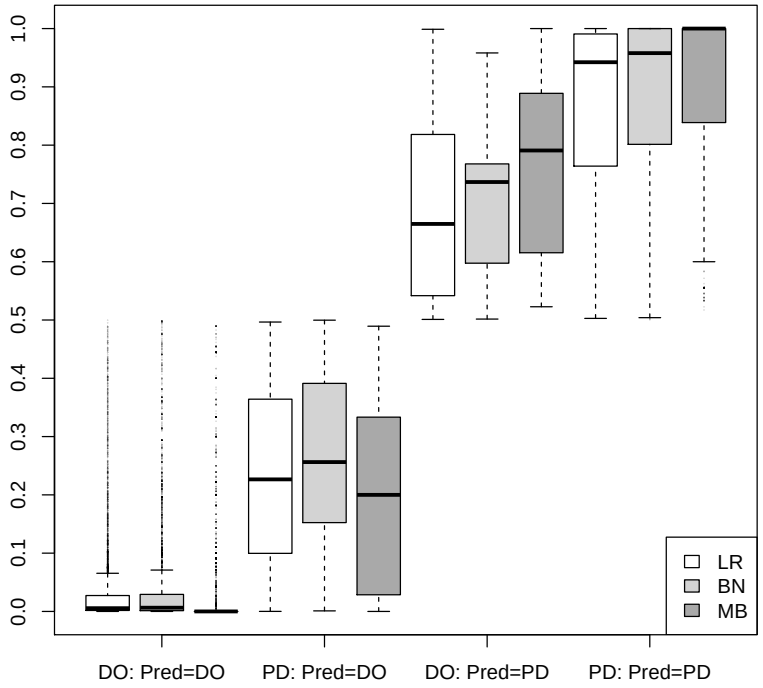

Figure 6.3: Boxplots of PD-likelihood scores for Logistic regression (LR), Bayesian Network (BN) and Memory-based learning (MB), sorted by the combination of the *actual* dative construction (DO, PD) and the *predicted* construction (Pred=DO, Pred=PD).

can be associated with cognitive processes: Bayesian Networks. In the third approach, we let go of the higher-level features and employed lexical items in a memory-based learning model.

Logistic regression is a statistical method that is convenient for several reasons: it is a multivariate approach, it is able to deal with non-numerical data, the models are fairly simple and logistic regression makes it possible to combine fixed variables (the features) and random variables (random effects). Also, this chapter confirms previous findings that logistic regression models with higher-level features are very powerful: 93.2% of the instances were classified correctly in 10-fold cross-validation, compared to a Verb/LenDif baseline of 89.3%. But there are also some drawbacks. First, it is often difficult to interpret the model coefficients because of the correlation between the input features. The regression model in this chapter showed that the collinearity in the data did not seem to have an effect on the significance and the sign of the regression coefficients for correlated features when the data set is large

(>11,000 instances). However, interpreting the actual values of the coefficients is not straightforward because it is unclear to what extent they are influenced by the collinearity. Second, it is difficult to link the regression models to cognitive processes, which receive increasingly more attention in linguistic research.

This motivated our choice for a second approach: a Bayesian Network that exploits the same higher-level features, and of which the graphical structure was based on theoretical reasoning. The network was equally accurate at the classification task as the logistic regression model: 93.2% in 10-fold cross-validation. The major advantages of the Bayesian Network approach are that it enables the modelling of the dependencies between the features explicitly, that it allows introducing hidden nodes that summarise other nodes, and that Bayesian inference can be associated with cognitive processes (Chater et al., 2006). Not only the classification accuracies, but also the classification of the individual constructions was similar to that by logistic regression: the Pearson correlation of the PD-likelihood scores was very high (0.95), and the classes based on these scores differed for only 3.3%: 384 (241 + 143) of the 11,784 instances. For research on alternations in general, a positive aspect of Bayesian Networks is that the number of feature values per node is not limited to two. This means that it is much easier to treat multi-class problems (such as the placement of adverbs in a sentence) than with logistic regression, which allows only pairwise comparisons. One of the risks of Bayesian Networks is that they may introduce circular reasoning. The topology of our networks was based on pre-existing theory and the outcomes of previous experiments. Today, there are no efficient techniques for learning the topology from the data; neither is it easy to determine whether arrows in a network that are mainly responsible for high classification accuracy indeed reflect the underlying cognitive processes. Also, the features on which the networks operate are derived from pre-existing theory. On the other hand, Bayesian Networks can help to falsify existing theories by showing that they cannot explain real (observed) language behaviour.

The accuracy of the memory-based learning approach, making use of lexical items instead of the higher-level features, did not differ significantly from the two approaches making use of higher-level features; in 10-fold cross-validation, the accuracy was 92.5%. However, the classification of the individual cases by the memory-based model differed most from that by the other two approaches, as we saw from the confusion matrices of the classifications and the Pearson correlations of the PD-likelihood scores. The instances that received a different class in memory-based learning than in the logistic regression model and the Bayesian Network were mostly instances with objects with large variation in words. Apparently, for cases where the possible words in the recipient or theme form a small set (e.g. in cases where it is a pronoun), the classifications

are similar to that by the other two models, but for the more unique objects, there are differences. In Section 6.4.3, we already mentioned that with the current data, it is impossible to say whether these differences are caused by the fact that memory-based learning models do not abstract away from the raw language input (while humans may do so), or by the fact that there is too little data available for the model to be able to classify correctly the less frequent cases (while humans receive many more exemplars). A model of human language acquisition in which language experience is stored and used in new situations, using general cognitive abilities instead of an innate language faculty (as for instance suggested in Daelemans & van den Bosch, 2005; Gahl & Yu, 2006; Bod, 2009), could therefore still be a suitable model for the dative alternation.

Regardless of the type of input features and the type of modelling technique, the largest part of the instances (92.0%) received the same class when training and testing on the same data, most of which (89.2% of all instances) were classified correctly. Seeing that the baseline using only the verb and the length difference already scores an accuracy of 89.6%, and all three approaches used these two features, this high level of agreement is not surprising. Also, we should note that several types of – often somewhat complex – dative constructions (e.g. passive and imperative clauses, clausal objects, etc.) were filtered out in our semi-automatic data collection. The filtering was partly the result of our decision to prevent the influence of other types of syntactic variation (passive versus active voice, declarative versus interrogative mode, the placement of adverbials, etc.). For the other part, they were an artifact of the approach chosen: keeping only those instances for which the higher-level feature values could be established, those that contained a verb in our list of dative verbs, and those that could be detected by the syntactic parser employed. As a result, only the more prototypical instances of the dative alternation are taken into account in this chapter. It is unclear how including the more complex constructions would have affected the predictive power of the different models considered and the explanatory value of the different higher-level features. Quite likely, including phrasal objects would have complicated the annotation for the higher level features and the feature selection in memory-based learning. Also, it is quite possible that the identity of the verb and the length of the objects are less predictive in the more complex constructions.

Nonetheless, the three full models provide significantly more accurate predictions than the baseline using verb and length difference only. Both the higher-level features and the lexical features may thus play a role in choosing one of the dative constructions. Seeing the small improvement over the baseline, however, it seems that in the data set used, the role of the features is limited and therefore difficult to establish. For now, this means that we cannot

be certain that humans make use of abstract semantic properties such as animacy and definiteness when choosing between the two dative constructions. At the same time, it appears that different verbs come with their own preferred constructions, which might give credibility to a theory based on memory-based processes. Also, one may speculate that realising the shortest (and usually given) object first frees memory and processing capacity for articulating the longer (and usually new) one, especially in spontaneous speech.

For the time being, we cannot draw hard and fast conclusions about which modelling technique is best suited to our purposes. Instead of only focussing on the static representations of already produced language (corpus data) as done in this thesis, research should also be directed at the exploration of models and feature representations that can be more closely linked to cognitive processes in *online* language production. Also, the studies should be extended to other syntactic alternations and other languages, to see how the feature representations and models hold across syntactic constructions and across languages.

7

Summary and conclusion

In this thesis, we addressed various choices that linguists must make when studying the dative alternation, as introduced in Chapter 1: which features to include in the study (*variable selection*), how to define and annotate the features used (*feature definition*), how to obtain an annotated data set that is sufficiently large (*(automatic) data collection*), how to study the alternation across different speaker groups (*comparison of speaker groups*) and how to interpret models found with various techniques (*model interpretation*). In this chapter, we summarise our findings for these five methodological issues addressed in the core chapters. We end this chapter with a general conclusion of the thesis, and provide suggestions for future research.

7.1 Summary of the findings

Variable selection

In Chapter 2, we addressed the research question:

Is it justified to report only one 'optimal' regression model, if models can be built in several different ways?

We built regular and mixed (i.e. containing a random effect) logistic regression models in order to explain the British English dative alternation, using a data set of 930 instances taken from the ICE-GB Corpus, manually annotated for the features introduced in Chapter 1 (the Ice-Trad data set). The regular and the mixed models were constructed following three different approaches: (1) providing the algorithms with all variables and keeping the significant ones, (2) starting with an empty model and successively adding the most predictive variables, and (3) starting with a model with all features and successively removing the least predictive variables. The six models showed some overlap in the variables that were regarded significant. These variables showed the

same effects as found by Bresnan et al. (2007): pronominal, relatively short, local (first or second person), discourse given, definite and concrete objects typically preceded objects with the opposite characteristics. Four variables were selected as interactions with medium, but only one of them was selected in more than one model: the effect of the discourse givenness of the recipient seemed to be stronger in written than in spoken language.

The models fitted the data better when verb sense was included as a random effect. The six methods we applied led to six different selections of variables and thus to six different models. We argued that this effect may have been due to the relatively small size of the data set used, which was one of the motivations for the automatic data extraction presented in Chapter 4. In a linguistic study using a small or medium-sized data set such as ours, it is not very clear how to select the model that fits the research goal best (combining the features suggested in the literature and test the combination on real data). Also, we prefer a model that is interpretable in the framework of some linguistic theory and that, ideally, reflects the processes in human brains. It is uncertain how (and if) we can evaluate a model in this sense. We therefore concluded that linguists should be careful in choosing a single variable selection approach and drawing conclusions from one model only; similar models should first be obtained with a number of different (types of) data sets.

Feature definition: concreteness

Chapter 3 focussed on the definition of the feature 'concreteness', aiming at answering the question:

What is the impact of different instantiations of the definition of the feature 'concreteness' on the actual labels given to corpus data, and on the outcome of syntactic research using this data?

We compared approaches to establish the concreteness of nouns varying in the definition used, in the scale of the values that could be assigned (interval, ordinal, nominal), the noun level they took as basis (token, sense, type) and the manner in which the values were assigned (manually, automatically, semi-automatically).

With respect to the impact on the actual labels given to corpus data, we found that the concreteness labels assigned to 68,848 nouns in the SemCor Corpus (the SemCor data set) showed considerable variation across the four labelling approaches we employed. The labellings following the definition of 'specificity' instead of 'sensory perceivability', differed most from the others. The fact that two approaches (one using a bootstrapping approach and one employing the MRC Psycholinguistic Database) mostly classified noun types, ignoring the word sense and the context, seemed to have no effect. We also

failed to find an effect for the measurement scale used and the manner of annotation.

In the second part of the chapter, we used 619 instances from the ICE-TRAD data set (the DATIVE data set) to build several regression models predicting the construction used, each using a different type of concreteness as a feature. The effects of the different types of concreteness varied considerably. Concreteness defined as 'specificity' did not seem to play a role in the dative alternation. When defined as 'sensory perceivability', concreteness only seemed to play a role when the approach included manual input. Again, we saw that the noun level and the measurement scale used have no clear effect, although the most significant regression coefficient was found for the only true token-based approach in the present research: the manual approach defined in the annotation manual in the Appendix.

To investigate to what degree humans agree about the concreteness of words in context, we employed a crowdsourcing experiment in which we asked workers to rate the concreteness of nouns presented in context. The human ratings showed that (also) for humans, there were instances that were clearly concrete or abstract, but also many instances for which humans disagreed about the concreteness. In cases where the words were (relatively) concrete in the definition of 'sensory perceivability', and relatively abstract in that of 'specificity', people seemed to focus most on the definition of 'sensory perceivability'.

We concluded that results concerning the concreteness in syntactic research can only be interpreted when taking into account the annotation scheme used (especially with respect to the definition used and the presence of human intervention) and the type of data that is being analysed (mostly because of the coverage issues of the resources we employed, and the differences in the concreteness ratings of individual language users).

Automatic data collection

In Chapter 4, we presented and evaluated an approach for automatically obtaining a data set for studying the dative alternation, with the research question:

Is data that is obtained and annotated automatically suitable for linguistic research, even if the data may contain a certain proportion of errors?

To address this question, we compared data that was extracted and annotated (semi-)automatically to two data sets that were manually obtained: 930 instances collected in Chapter 2 from the ICE-GB corpus of spoken and written British English (ICE-TRAD, used as development data), and 2,349 instances collected by Bresnan et al. (2007) from the Switchboard corpus of spoken American

English (Swb-Trad, used as test data).

In this thesis, we decided to use an off-the-shelf syntactic parser that distinguishes both dative constructions explicitly: the Functional Dependency Grammar (FDG) parser developed at Connexor (Tapanainen & Järvinen, 1997). However, the FDG parser was not very successful in detecting instances of the dative alternation: in combination with our two filtering modules, the recall was 66.6% for the ICE-GB (Ice-Auto, various types of spoken and written data) and 55.0% for Switchboard (Swb-Auto, spontaneous speech only), and the precision was 71.2% for Ice-Auto and only 48.0% for Swb-Auto. When we used the various data sets to build regression models, we found that the model for Ice-Auto contained four significant effects that were not found for Ice-Trad. The analysis of the errors in Ice-Auto showed that the FDG parser had most difficulty with spoken material, with longer sentences and with PP-attachment. We concluded that we needed one (simple) manual step: manually checking the relevance of the candidates that were found automatically, after which the approved instances could be annotated automatically. The models built on only the instances that were manually approved (Ice-Semi and Swb-Semi) appeared to be very similar to those found for Ice-Trad and Swb-Trad. We concluded that sensible data sets can be obtained even with an off-the-shelf parser that yields a low recall of recognising dative constructions (55% to 66.6%).

With respect to the second step, feature extraction, we concluded that our rather straightforward feature extraction algorithm was suitable for automatically annotating the instances with the features suggested in the literature (as introduced in Table 1.1 of Chapter 1 and used throughout this thesis).[1] The κ scores between the manual and the automatic annotations were similar to scores found between human annotators, except for the intuitively most difficult features: animacy, concreteness and discourse givenness. Only the automatic annotation of the concreteness of theme was so dissimilar from the human annotations that it notably influenced the regression models. When excluding this feature, the models built on Ice-Semi and Swb-Semi (with automatic annotations) were very similar to the ones obtained for Ice-Trad and Swb-Trad (with manual annotations). The differences between the models did not seem to be caused by the errors in the automatic annotations, but by properties inherent to the data sets: multiple correlations between the features, and the use of different definitions for the same feature.

In the discussion in Section 4.6, we mentioned that the definitions of the features based on Bresnan et al. (2007) are by no means definitive. Even after following their definitions as strictly as possible, there were differences

[1] The feature extraction script can be downloaded from `http://daphnetheijssen.ruhosting.nl/downloads`.

in the annotations made. Furthermore, two different data sets, though drawn from the same population (i.e. the English language as represented in the ICE-GB and Switchboard corpora), can result in different models because their composition differs. As a result, the features that showed no significance in our models could still play a role in another data set. Since the influence of the composition of a data set usually decreases when data sets become larger, we applied the semi-automatic approach to the 100-million word BNC (BNC Consortium, 2007) in Chapter 6.

Comparison of speaker groups

Chapter 5 presented a study that merged corpus linguistics, psycholinguistics and sociolinguistics, in which we treated the question:

What are the differences and/or similarities in the dative alternation of British, American and Australian language users varying in age and gender?

We treated this question with the help of a corpus study (as used in the preceding chapters) and a judgement study in which we asked participants to divide 100 points over the two alternative constructions presented in context. The more points given, the more natural the construction sounded to the participants. Both studies showed that there are certain patterns in the data sets that are in line with each other and with previous work (see also Chapters 2 and 4): animate usually precedes inanimate, definite usually precedes indefinite, shorter usually precedes longer and pronouns usually precedes nonpronouns.

American English and Australian English originated from British English at different times in history. Despite the cultural exchange that dominates our modern society, we found other distributional differences between British and American English than between British and Australian English. In our corpus study, we found that in contemporary spontaneous spoken English, American speakers show a stronger tendency towards the double object construction when the theme is indefinite (e.g. *give him a book*) than British speakers, but a less strong preference for the prepositional dative as the recipient length increases relative to the theme. We suggested that the relative importance of these two features may have evolved further in American English than in British English, following the developments found in the diachronic study by Wolk et al. (2012), but the effects were not confirmed in our judgement study. The judgement study indicated that the effect of length difference is stronger for Australian participants than for British participants, as also found by Bresnan and Ford (2010).

We also investigated the differences and similarities in the dative alternation made by participants varying in *age and gender*. In many dialects in

British English, it is common to use double object constructions with pronominal themes, e.g. *give the man it* (Siewierska & Hollmann, 2007; Haddican, 2010). Our study showed that the younger British participants are more in favour of the prepositional dative variant when the theme is a pronoun (*give it to the man*) than the older participants, thus moving away from the dialectal construction. The US judgements showed that, in general, the prepositional dative construction is most popular with the younger participants, and the same effect was found for Australian men. Also, regardless of age, Australian men were more positive towards using the prepositional dative than Australian women, which supported the findings in Bresnan and Ford (2010).

The results, and comparisons to existing research, could not always be interpreted straightforwardly since the underlying data sets were not fully compatible with respect to the genres included, the annotations for the features and the items selected. Also, we found that most of the variance in the choice between the two constructions could be explained by the random effects: the verb and the theme head (in the corpus study), or the verb and the participant (in the judgement study). This means that the features under investigation played a significant, but minor, role. The predictive power of individual speakers and test items is often found in (psycho)linguistic studies (cf. Baayen, 2008). We concluded that in order to establish the role of the features in cognitive processes, on top of the effect of frequent lexico-syntactic patterns and participant-specific preferences, future research should be directed at studying the dative alternation and other syntactic alternations in languages other than English.

Model interpretation

In Chapter 6, we focussed on the following research question:

How suitable are regression models, memory-based learning and Bayesian networks for studying the dative alternation?

Regression models combining higher-level features have been used throughout this thesis. Regression is a versatile statistical method: it is a multivariate approach, it can deal with non-numerical data, the models are fairly simple and it allows to combine fixed variables (the features) and random variables (random effects). In this chapter, we applied the semi-automatic data extraction approach from Chapter 4 to the British National Corpus (BNC Consortium, 2007), resulting in a data set of 11,784 instances. We used this set to build a regression model, which again showed that the higher level features are very predictive: 93.2% of the instances were classified correctly in 10-fold cross-validation, significantly better than the Verb/LenDif baseline of 89.3%. The regression model showed that the collinearity in the data had no effect on

the *significance* of the correlated features, since the data set used was large (11,784 instances). Collinearity can also cause coefficients to *flip sign*: if two features are (strongly) correlated, the feature with the highest correlation with the dative construction may leave only a residual to explain by the other feature, which may then receive a coefficient with the opposite sign. But seeing that the patterns found were consistent with those found in the vast body of research (including studies using experimental data, and studies investigating the features one at a time), the influence of the collinearity did not seem very large. Some drawbacks of the regression approach were that interpreting the coefficient values was not straightforward, and it was difficult to link the regression models to cognitive processes.

We used the same high-level features in a Bayesian Network, a modelling technique that can be associated with cognitive processes (Chater et al., 2006). It also reached 93.2% classification accuracy in 10-fold cross-validation, and the classifications were similar to those by logistic regression (Pearson correlation of 0.95, only 384 of the 11,784 instances differed). Advantages of the Bayesian Network approach were that it makes it possible to model the dependencies between the features explicitly and that it allows introducing *hidden* nodes that summarise the combined effects of several features. Also, for research on alternations in general, a positive aspect is that since the number of feature values per node is not limited to two, it is easy to treat multi-class problems such as the placement of adverbs in a sentence. A risk of Bayesian Networks is that they may introduce circular reasoning: the high-level features and the topology of our network were based on pre-existing theory and the outcomes of previous experiments. Today, there are no efficient techniques for learning the topology from the data. Moreover, it is not easy to verify that the links in a network yielding high classification accuracy do indeed reflect the underlying cognitive processes.

For the memory-based learning approach, we employed surface features only, assuming a model of human language acquisition in which language experience is stored and used in new situations, using general cognitive abilities instead of an innate language faculty (as for example suggested in Daelemans & van den Bosch, 2005; Gahl & Yu, 2006; Bod, 2009). The prediction accuracy reached (92.5% in 10-fold cross-validation) did not differ significantly from the two approaches using higher-level features. However, the individual classifications differed substantially from those by the other two approaches, especially for instances with more uniquely expressed themes or recipients (e.g. full noun phrases instead of pronouns). With the research in this chapter, we could not establish whether these differences were caused by the fact that memory-based learning models do not abstract away from the raw language input (while humans may do so), or by the fact that there is too little data

available for the model to be able to classify correctly the less frequent cases (while humans receive many more exemplars).

The baseline using only the verb and the length difference already yielded a classification accuracy of 89.6%, but the three full models provided significantly more accurate predictions than this baseline. Thus, it seems that both the higher level features and the lexical features may play a role in choosing one of the dative constructions. So for now, we cannot be certain that humans make use of abstract semantic properties such as animacy and definiteness when choosing between the two dative constructions. It does seem that different verbs come with their own preferred constructions, which might give credibility to a theory based on memory-based processes. Also, one may speculate that realising the shortest (and usually given) object first frees memory and processing capacity for articulating the longer (and often newly introduced) one, especially in spontaneous speech.

7.2 General conclusion and suggestions for future research

The previous section summarised our conclusions with respect to the five research questions of this thesis. There are also some general observations to be made.

The effect of the methodology adopted

We discovered that some choices made in the methodology can substantially influence the results. Chapter 2 showed that variable selection has a great impact on the logistic regression models found. In Chapter 3, we saw that the same is true for the definition and annotation procedure used for the predictive features used (in this case: the concreteness of the theme). Also, we had to conclude that the commonly used logistic regression approach not only has positive aspects, but also some drawbacks. Most importantly, for smaller data sets, the collinearity of the features may influence their significance in the models (Chapters 2, 3, 4).

Seeing these influences of methodological choices on the eventual results, our recommendation is to be careful when drawing conclusions from the models found. To be more certain about the findings, the same results should be obtained with different models and/or different data sets. A good way to achieve this is to compare the findings to those found by other researchers. To enable such comparison, however, it is crucial that researchers report the exact methodology and data used to arrive at the results.

Difficulties with the data

Throughout this thesis, we came across some difficulties with respect to the data used. For instance, we discovered that even when following the annotation instructions provided by other researchers, there may be relevant differences between the labellings (Chapter 4). In Chapters 3 and 5, we demonstrated that also in corpus-driven studies, it is helpful to include data from questionnaires or other experimental studies as well.

With any type of data, it is often difficult to collect enough data to prevent sparseness problems in models, such as the undesirable influence of the variable selection method used to build the models (Chapter 2). In Chapter 4, we presented an approach to solve the inconsistency and sparseness of the data by extracting dative instances semi-automatically from existing corpora. For this approach, the corpora needed no annotations; raw text only was sufficient. We successfully applied this approach to the BNC in Chapter 6, and we believe this is a promising result for corpus linguists studying syntactic alternations. We should note, however, that off-the-shelf tools such as syntactic parsers are usually tailored towards clean, (more often than not edited) written language. As a result, we found that the recall of recognising dative constructions was especially low for spontaneous speech (only 55.0% for Switchboard, see Chapter 4). Even though our evaluations did not reveal it, there is a risk that the cases found and the cases missed have systematic properties that could have affected the models. If this is the case, the use of off-the-shelf tools may not be sufficient to extract a data set that is representative of the dative alternation, especially in spontaneous speech.

Linguistic findings

We found some results that are relevant for linguistic research on the dative alternation. Across the chapters, the higher-level features introduced in Chapter 1 confirmed the patterns already established in previous research (e.g. Bresnan et al., 2007):

animate usually preceded inanimate
definite usually preceded indefinite
given usually preceded nongiven
shorter usually preceded longer
pronoun usually preceded nonpronoun

These patterns have also been found for other syntactic alternations (e.g. also in the genitive alternation, Szmrecsányi, 2010), varieties of English (e.g. British, American, Australian, New Zealand, Indian and African-American English), and types of data (speaking versus writing, corpus or experimental).

We also found some subtle differences between the relative influences of the features in different types of data, e.g. between speech and writing (Chapters 2, 4 and 6), between different speaker groups (Chapter 5) and between productions and judgements (Chapter 5). Given the effect that the methodology of data extraction and analysis has on the models found, we are hesitant to draw firm conclusions. We believe that more research is needed to attest the effect that these extralinguistic factors have on the dative alternation. As mentioned above, however, comparing research findings will only be feasible if researchers report their exact research procedure.

The future of modelling syntactic alternation

In the models presented in this thesis, we saw that most of the variance in the models was explained by the lengths of the objects, the verb and/or the participant. This made us question the role and the justification of the higher-level features in cognitive models of syntactic alternation.

Also, given the fact that language is an aspect of cognition, it may be a good idea to move away from the use of regression models, and towards modelling techniques that are associated with cognitive processes such as those in Chapter 6. It may thus be time for the common approach of employing higher-level features in a regression model to make room for new feature sets and new modelling techniques. But to truly establish the universality of any feature set and any model, research should also be directed at other alternations and other languages.

Bibliography

Agirre, E., Baldwin, T., & Martinez, D. (2008). Improving parsing and PP attachment performance with sense information. In *Proceedings of the Workshop on Human Language Technologies at the 46th annual meeting of the Association for Computational Linguistics (ACL 2008)* (pp. 317–325).

Anagnostopoulou, E. (2005). Cross-linguistic and cross-categorial variation of datives. In M. Stavrou & A. Terzi (Eds.), *Advances in Greek generative grammar* (pp. 61–126). Amsterdam, The Netherlands: John Benjamins.

Arnold, J., Wasow, T., Losongco, A., & Ginstrom, R. (2000). Heaviness vs. newness: The effects of complexity and information structure on constituent ordering. *Language*, *76*(1), 28–55.

Artstein, R., & Poesio, M. (2008). An empirically based system for processing definite descriptions. *Computational Linguistics*, *34*(4), 555–596.

Baayen, R. H. (2008). *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen, R. H. (2011). Corpus linguistics and naive discriminative learning. *Brazilian Journal of Applied Linguistics*, *11*, 295–328.

Babanoğlu, M. P. (2007). *The acquisition of English dative alternation by Turkish adult learners of English*. Unpublished master's thesis, Department of English language teaching, Çukurova University.

Baker, K., & Brew, C. (2008). Multilingual animacy classification by sparse logistic regression. *Ohio State Working Papers in Linguistics*.

Bates, D. (2005). Fitting linear mixed models in R. *R News*, *5*(1), 27–30.

Bean, D. L., & Riloff, E. (1999). Corpus-based identification of non-anaphoric noun phrases. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics (ACL 1999)* (pp. 373–380).

Beavers, J., & Nishida, C. (2010). The Spanish dative alternation revisited. In S. Colina, A. Olarrea, & A. Carvalho (Eds.), *Romance Linguistics 2009: Selected papers from the 39th Linguistic Symposium of Romance Languages* (pp. 217–230).

Behaghel, O. (1909). Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen*, *25*, 110–142.

Biber, D. (1985). Investigating macroscopic textual variation through multi feature/multi dimensional analyses. *Linguistics*, *23*(2), 337–360.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D., Finegan, E., & Atkinson, D. (1994). ARCHER and its challenges: compiling and exploring A Representative Corpus of Historical English Registers. In U. Fries, G. Tottie, & P. Schneider (Eds.), *Creating and using English language corpora: Papers from the fourteenth International Conference on English Language Research and Computerized Corpora* (pp. 1–13).

Bikel, D. M. (2002). Design of a multi-lingual, parallel-processing statistical parsing engine. In *Proceedings of the Human Language Technology conference 2002 (HLT-2002).*

Blackwell, A. A. (2005). Acquiring the English adjective lexicon: relationships with input properties and adjectival semantic typology. *Child Language*, *32*(3), 535–562.

BNC Consortium. (2007). The British National Corpus, version 3 (BNC XML Edition) [Computer software manual]. Available from `http://www.natcorp.ox.ac.uk/`

Bock, K., & Irwin, D. (1980). Syntactic effects of information availability in sentence production. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 467–484.

Bock, K., Loebell, H., & Morey, R. (1992). From conceptual roles to structural relations: Bridging the syntactic cleft. *Psychological Review*, *99*(1), 150–171.

Bod, R. (2007). Is the end of supervised parsing in sight? In *Proceedings of the 45th annual meeting of the Association for Computational Linguistics (ACL 2007)* (pp. 400–407).

Bod, R. (2009). From exemplar to grammar: A probabilistic analogy-based model of language learning. *Cognitive Science*, *33*(5), 752–793.

Bod, R., Hay, J., & Jannedy, S. (2003). *Probabilistic Linguistics*. Cambridge, MA, USA: MIT Press.

Bolinger, D. (1977). *Meaning and form*. London, UK: Longman.

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). Predicting the dative alternation. In G. Bouma, I. Kraemer, & J. Zwarts (Eds.), *Cognitive Foundations of Interpretation* (pp. 69–94). Amsterdam, The Netherlands: Royal Netherlands Academy of Science.

Bresnan, J., & Ford, M. (2010). Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language*, *86*(1), 168–213.

Bresnan, J., & Hay, J. (2008). Gradient grammar: An effect of animacy on the syntax of *give* in New Zealand and American English. *Lingua*, *118*(2), 245–259.

Bresnan, J., & Nikitina, T. (2009). The gradience of the dative alternation. In L. Uyechi & L.-H. Wee (Eds.), *Reality exploration and discovery: Pattern*

interaction in language and life (pp. 161–184). Stanford, CA, USA: CSLI Publications.

Chang, C.-C., & Lin, C.-J. (2001). LIBSVM: a library for support vector machines [Computer software manual]. Available from `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

Changizi, M. A. (2008). Economically organized hierarchies in WordNet and the Oxford English Dictionary. *Cognitive Systems Research*, *9*(3), 214–228.

Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2010). Bayesian models of cognition. *WIREs Cognitive Science*, *1*, 811 – 823.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*(7), 287–291. (Introduction to the special issue on Probabilistic models of cognition.)

Choi, H.-W. (2007). Length and order: A corpus study of Korean dative-accusative construction. *Discourse and Cognition*, *14*(3), 207–227.

Colleman, T. (2006). *De Nederlandse datiefalternantie: een constructioneel en corpusgebaseerd onderzoek [The Dutch dative alternation: a constructional and corpus-based study]*. Unpublished doctoral dissertation, Ghent University.

Collins, P. (1995). The indirect object construction in English: an informational approach. *Linguistics*, *33*(1), 35–49.

Coltheart, M. (1981). *MRC Psycholinguistic Database user manual: Version 1*. London, UK: Birkbeck College.

Daelemans, W., & van den Bosch, A. (2005). *Memory-Based Language Processing*. Cambridge, UK: Cambridge University Press.

Daelemans, W., Zavrel, J., van der Sloot, K., & van den Bosch, A. (2010). TiMBL: Tilburg Memory Based Learner version 6.3 Reference Guide [Computer software manual].

Daudaravičius, V., & Marcinkevičiene, R. (2004). Gravity counts for the boundaries of collocations. *International Journal of Corpus Linguistics*, *9*(2), 321–348.

Davidse, K. (1996). Functional dimensions of the dative in English. In W. van Belle & W. van Langendonck (Eds.), *The dative, vol. 1: Descriptive studies* (pp. 289–338). Amsterdam, The Netherlands: John Benjamins.

de Marneffe, M.-C., Grimm, S., Arnon, I., Kirby, S., & Bresnan, J. (2012). A statistical model of the grammatical choices in child production of dative sentences. *Language and Cognitive Processes*, *21*, 25 – 61.

Dowman, M. (2004). *Colour terms, syntax and Bayes: Modelling acquisition and evolution*. Unpublished doctoral dissertation, School of Information Technologies, University of Sydney.

Estival, D., & Myhill, J. (1988). Formal and functional aspects of the development

from passive to ergative systems. In M. Shibatani (Ed.), *Passive and voice* (pp. 441–491). Amsterdam, The Netherlands: John Benjamins.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA, USA: MIT Press.

Gahl, S., & Yu, A. C. (2006). Introduction to the special issue on exemplar-based models in linguistics. *The Linguistic Review*, *23*(3), 213–216.

Garretson, G. (2003). *Coding manual for the project "Optimal typology of determiner phrases".* (Unpublished manuscript, Boston University)

Geeraerts, D., Kristiansen, G., & Peirsman, Y. (2010). *Advances in Cognitive Sociolinguistics*. Berlin, Germany: Walter de Gruyter.

Girju, R., Roth, D., & Sammons, M. (2005). Token-level disambiguation of VerbNet classes. In K. Erk, A. Melinger, & S. Schulte im Walde (Eds.), *Proceedings of the interdisciplinary workshop on the Identification and Representation of Verb Features and Verb Classes* (pp. 56–61).

Givón, T. (1984). Direct object and dative shifting: Semantic and pragmatic case. In F. Plank (Ed.), *Objects: Towards a theory of grammatical relations* (pp. 151–182). New York, NY, USA: Academic Press.

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *Proceedings of ICASSP-92* (pp. 517–520).

Gomes, C. A. (2003). Dative alternation in Brazilian Portuguese: typology and constraints. *Language Design: Journal of Theoretical and Experimental Linguistics*, *5*, 67–78.

Grafmiller, J. (2012). Variation in English genitives across modality and genre. *English Language and Linguistics*. (Forthcoming.)

Greenbaum, S. (Ed.). (1996). *Comparing English worldwide: The International Corpus of English*. Oxford, UK: Clarendon Press.

Gregory, M. (1967). Aspects of varieties differentiation. *Journal of Linguistics*, *3*(2), 177–197.

Gries, S. T. (2003). Towards a corpus-based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics*, *1*(4), 1–27.

Gries, S. T. (2010). Useful statistics for corpus linguistics. In A. Sánchez & M. Almela (Eds.), *A mosaic of corpus linguistics: selected approaches* (pp. 269–291). Frankfurt am Main, Germany: Peter Lang.

Gries, S. T., & Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics*, *9*(1), 97–129.

Grimm, S., & Bresnan, J. (2009). Spatiotemporal variation in the dative alternation: a study of four corpora of British and American English. In *Third International Conference Grammar and Corpora*.

Grondelaers, S., & Speelman, D. (2007). A variationist account of constituent ordering in presentative sentences in Belgian Dutch. *Corpus Linguistics and Linguistic Theory*, *3*(2), 161–193.

Haddican, W. (2010). Theme-goal ditransitives and theme passivisation in British English dialects. *Lingua*, *120*(10), 2424–2443.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations*, *11*(1), 10–18.

Harris, Z. (1954). Distributional structure. *Word*, *10*(23), 146–162.

Hinrichs, L., Smith, N., & Waibel, B. (2007). The part-of-speech-tagged 'Brown' corpora: A manual of information, including pointers for successful use [Computer software manual].

Hinrichs, L., & Szmrecsányi, B. (2007). Recent changes in the function and frequency of Standard English genitive constructions: a multivariate analysis of tagged corpora. *English Language and Linguistics*, *11*(3), 437–474.

Ide, N., & Romary, L. (2008). Towards international standards for language resources. In L. Dybkjær, W. Minker, & H. Hemsen (Eds.), *Evaluation of text and speech systems* (pp. 69–94). Springer.

Izenman, A. J. (2008). *Modern multivariate statistical techniques: Regression, classification, and manifold learning*. New York, NY, USA: Springer.

Jankowski, B. (2009). Grammatical and register variation and change: A multi-corpora perspective on the English genitive. In *American Association for Corpus Linguistics (AACL 2009).*

Joanis, E., Stevenson, S., & James, D. (2008). A general feature space for automatic verb classification. *Natural Language Engineering*, *14*(3), 337–367.

Kabadjov, M. A. (2007). *A comprehensive evaluation of anaphora resolution and discourse-new classification*. Unpublished doctoral dissertation, Department of Computer Science, University of Essex.

Keller, F., Corley, M., Corley, S., Crocker, M. W., & Trewin, S. (1999). Gsearch: A tool for syntactic investigation of unparsed corpora. In *Proceedings of the EACL Workshop on linguistically interpreted corpora* (pp. 56–63).

Kendall, T., Bresnan, J., & van Herk, G. (2011). The dative alternation in African American English: Researching syntactic variation and change across sociolinguistic datasets. *Corpus Linguistics and Linguistic Theory*, *7*(2), 229–244.

Kipper, K., Dang, H. T., & Palmer, M. (2000). Class-based construction of a verb lexicon. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000)* (pp. 691–696).

Koiter, J. (2006). *Visualizing Inference in Bayesian Networks*. Unpublished

master's thesis, Man-machine interaction group, Delft University of Technology.

Korhonen, A. (2009). Automatic lexical classification - balancing between machine learning and linguistics. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation* (pp. 19–28).

Kristiansen, G., & Dirven, R. (2008). *Cognitive Sociolinguistics: Language variation, cultural models, social systems*. Berlin, Germany: Mouton de Gruyter.

Kübler, S. (2007). How do treebank annotation schemes influence parsing results? Or how not to compare apples and oranges. In N. Nicolov, K. Boncheva, G. Angelova, & R. Mitkov (Eds.), *Recent advances in Natural Language Processing IV: Selected papers from RANLP 2005* (pp. 79–88). Amsterdam, The Netherlands: John Benjamins.

Kuijjer, C. (2007). *Semantic lexicon expansion using bootstrapping and syntax-based, contextual extraction patterns*. Unpublished master's thesis, Information Sciences, University of Amsterdam.

Lapata, M. (1999). Acquiring lexical generalizations from corpora: a case study for diathesis alternations. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics (ACL 1999)* (pp. 397–404).

Lapata, M., & Brew, C. (2004). Verb Class Disambiguation Using Informative Priors. *Computational Linguistics*, *30*(1), 45–73.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. Chicago, IL, USA: The University of Chicago.

Levin, B. (2008). Dative verbs: A crosslinguistic perspective. *Lingvisticæ Investigationes*, *31*(2), 285–312.

Li, J., & Brew, C. (2008). Which are the best features for automatic verb classification. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics* (pp. 434–442).

Lyons, J. (1977). *Semantics* (Vol. 2). Cambridge, UK: Cambridge University Press.

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.

McCarthy, D. (2001). *Lexical acquisition at the syntax-semantics interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences*. Unpublished doctoral dissertation, University of Sussex.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., et al. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in cognitive sciences*, *14*, 348 – 356.

Miller, G. A., Leacock, C., Tengi, R., & Bunker, R. T. (1993). A semantic concor-

dance. In *Proceedings of the 3 DARPA Workshop on Human Language Technology* (pp. 303–308).

Miyagawa, S., & Tsujioka, T. (2004). Argument structure and ditransitive verbs in Japanese. *Journal of East Asian Linguistics*, *13*(1), 1–38.

Moore, R. K. (2003). A comparison of the data requirements of automatic speech recognition systems and human listeners. In *Proceedings of EUROSPEECH'03* (pp. 2582–2584).

Mukherjee, J., & Hoffmann, S. (2006). Describing verb-complementational profiles of New Englishes. *English World-Wide*, *27*(2), 147-173.

Nancarrow, O., & Atwell, E. (2007). A comparative study of the tagging of adverbs in modern English corpora. In *Proceedings of Corpus Linguistics 2007 (CL2007).*

Ng, V., & Cardie, C. (2002). Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002)* (pp. 730–736).

Oostdijk, N. (1996). Using the TOSCA analysis system to analyse a software manual corpus. In R. E. E. Sutcliffe, H.-D. Koch, & A. McElligott (Eds.), *Industrial parsing of software manuals* (pp. 179–206). Amsterdam, The Netherlands: Rodopi.

Orăsan, C., & Evans, R. (2007). NP animacy identification for anaphora resolution. *Journal of Artificial Intelligence Research*, *29*, 79–103.

Owen, A. B. (2007). Infinitely imbalanced logistic regression. *The Journal of Machine Learning Research*, *8*, 761–773.

Ozón, G. A. (2009). *Alternating ditransitives in English: A corpus-based study.* Unpublished doctoral dissertation, University College London.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of Plausible Inference*. Cambridge, MA, USA: Morgan Kaufmann Publishers, Inc.

Pickering, M. J., & Garrod, S. (2005). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, *11*(3), 105–110.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA, USA: MIT Press.

Poesio, M., Uryupina, O., Vieira, R., Kabadjov, M. A., & Goulart, R. (2004). Discourse-new detectors for definite description resolution: A survey and a preliminary proposal. In *Proceedings of the Workshop on Reference Resolution at the 42nd annual meeting of the Association for Computational Linguistics (ACL 2004)* (pp. 47–54).

Poibeau, T., & Messiant, C. (2008). Do we still need gold standards for evalu-

ation? In *Proceedings of the Language Resources and Evaluation Conference (LREC)* (pp. 547–552).

Prat-Sala, M., & Branigan, H. P. (2000). Discourse constraints on syntactic processing in language production: A cross-linguistic study in English and Spanish. *Journal of Memory and Language*, *42*(2), 168–182.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1972). *A grammar of contemporary English*. London, UK: Longman.

R Development Core Team. (2008). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from `http://www.R-project.org`

Rietveld, T., & van Hout, R. (2008). *Statistical techniques for the study of language and language behavior*. Berlin, Germany: Mouton de Gruyter.

Riloff, E., & Jones, R. (1999). Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)* (pp. 474–479).

Rooth, M., Riezler, S., Prescher, D., Carroll, G., & Beil, F. (1999). Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics (ACL 1999)* (pp. 104–111).

Rosenbach, A. (2003). Aspects of iconicity and economy in the choice between the *s*-genitive and the *of*-genitive in English. In G. Rohdenburg & B. Mondorf (Eds.), *Determinants of grammatical variation in English* (pp. 379–411). Berlin, Germany: Mouton de Gruyter.

Rosenbach, A. (2005). Animacy versus weight as determinants of grammatical variation in English. *Language*, *81*(3), 613–644.

Schmid, H.-J. (2000). *English abstract nouns as conceptual shells: From corpus to cognition*. Berlin, Germany: Mouton de Gruyter.

Schulte im Walde, S., Hying, C., Scheible, C., & Schmid, H. (2008). Combining EM Training and the MDL Principle for an Automatic Verb Classification incorporating Selectional Preferences. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics* (pp. 496–504).

Schulte Im Walde, S. (2009). The Induction of Verb Frames and Verb Classes from Corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An international handbook* (Vol. 2, pp. 952–972). Berlin, Germany: Mouton de Gruyter.

Sheather, S. J. (2009). *A modern approach to regression with R*. New York, NY, USA: Springer.

Shih, S., & Grafmiller, J. (2011). Weighing in on end weight. In *Annual Meeting of the Linguistic Society of America*.

Siewierska, A., & Hollmann, W. (2007). Ditransitive clauses in English with special reference to Lancashire dialect. In M. Hannay & G. J. van der

Steen (Eds.), *Structural-functional studies in English grammar: In honor of Lachlan Mackenzie* (pp. 83–102). Amsterdam, The Netherlands: John Benjamins.

Snyder, K. (2003). *The relationship between form and function in ditransitive constructions*. Unpublished doctoral dissertation, University of Pennsylvania, PA, USA.

Spreen, O., & Schulz, R. W. (1966). Parameters of abstraction, meaningfulness, and pronunciability for 329 nouns. *Journal of Verbal Learning and Verbal Behavior*, *5*, 459–468.

Sun, L., & Korhonen, A. (2009). Improving verb clustering with automatically acquired selectional preferences. In *Proceedings of the 2009 conference on Empirical Methods in Natural Language Processing (EMNLP 2009)* (pp. 638–647).

Szmrecsányi, B. (2005). Language users as creatures of habit: A corpus-based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory*, *1*(1), 113–149.

Szmrecsányi, B. (2006). *Morphosyntactic persistence in spoken English: A corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis*. Berlin, Germany: Mouton de Gruyter.

Szmrecsányi, B. (2010). The English genitive alternation in a cognitive sociolinguistics perspective. In D. Geeraerts, G. Kristiansen, & Y. Peirsman (Eds.), *Advances in Cognitive Sociolinguistics* (pp. 141–166). Berlin, Germany: Walter de Gruyter.

Szmrecsányi, B., & Hinrichs, L. (2008). Probabilistic determinants of genitive variation in spoken and written English: A multivariate comparison across time, space and genres. In T. Nevalainen, I. Taavitsainen, P. Pahta, & M. Korhonen (Eds.), *The dynamics of linguistic variation: Corpus evidence on English past and present* (pp. 291–309). Amsterdam, The Netherlands: John Benjamins.

Tagliamonte, S., & Jarmasz, L. (2008). Variation and change in the English genitive: A sociolinguistic perspective. In *Annual Meeting of the Linguistic Society of America*.

Tapanainen, P., & Järvinen, T. (1997). A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing* (pp. 64–71).

Theijssen, D. (2010). Variable selection in Logistic Regression: The British English dative alternation. In T. Icard & R. Muskens (Eds.), *Interfaces: Explorations in Logic, Language and Computation* (Vol. 6211 of Springer Lecture Notes in Artificial Intelligence, pp. 87–101).

Theijssen, D., Boves, L., van Halteren, H., & Oostdijk, N. (2011). Evaluating automatic annotation: Automatically detecting and enriching instances

of the dative alternation. *Language Resources and Evaluation*, DOI: 10.1007/s10579-011-9156-x.

Theijssen, D., Bresnan, J., Ford, M., & Boves, L. (2011). *In a land far far away... A probabilistic account of the dative alternation in British, American, and Australian English.* (Under review.)

Theijssen, D., ten Bosch, L., Boves, L., Cranen, B., & van Halteren, H. (2012). Choosing alternatives: Using Bayesian Networks and memory-based learning to study the dative alternation. *Corpus Linguistics and Linguistic Theory*. (Accepted for publication.)

Theijssen, D., van Halteren, H., Boves, L., & Oostdijk, N. (2011a). The more the merrier? How data set size and noisiness affect the accuracy of predicting the dative alternation. In *21st meeting of Computational Linguistics in the Netherlands (CLIN-21).* University College Ghent, Ghent, Belgium.

Theijssen, D., van Halteren, H., Boves, L., & Oostdijk, N. (2011b). On the difficulty of making concreteness concrete. *Computational Linguistics in the Netherlands Journal*, *1*, 61–77.

Theijssen, D., van Halteren, H., Fikkers, K., Groothoff, F., van Hoof, L., van de Sande, E., et al. (2009). A regression model for the English benefactive alternation. In B. Plank, E. Tjong Kim Sang, & T. Van de Cruys (Eds.), *Computational Linguistics in the Netherlands 2009: Selected papers from the nineteenth CLIN meeting* (pp. 115–130).

Theijssen, D., Verberne, S., Oostdijk, N., & Boves, L. (2007). Evaluating deep syntactic parsing: Using TOSCA for the analysis of *why*-questions. In P. Dirix, I. Schuurman, V. Vandeghinste, & F. Van Eynde (Eds.), *Computational Linguistics in the Netherlands 2006: Selected papers from the seventeenth CLIN meeting* (pp. 115–130).

Thelen, M., & Riloff, E. (2002). A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)* (pp. 214–221).

Thompson, S. A. (1990). Information flow and dative shift in English discourse. In J. A. Edmondson, C. Feagin, & P. Mühlhausler (Eds.), *Development and diversity: Language variation across time and space* (pp. 239–253). Dallas, Arlington, TX, USA: Summer Institute of Linguistics and University of Texas at Arlington.

Thompson, S. A. (1995). The iconicity of 'dative shift' in English: Considerations from information flow in discourse. In M. E. Landsberg (Ed.), *Syntactic iconicity and linguistic freezes* (pp. 155–175). Berlin, Germany: Mouton de Gruyter.

Uryupina, O. (2003). High-precision identification of discourse new and unique noun phrases. In *Proceedings of the Student Workshop at the 41st annual*

meeting of the Association for Computational Linguistics (ACL 2003) (pp. 80–86).

van den Bosch, A. (2004). Wrapped progressive sampling search for optimizing learning algorithm parameters. In R. Verbrugge, N. Taatgen, & L. Schomaker (Eds.), *Proceedings of the 16th Belgian-Dutch Conference on Artificial Intelligence.*

Vieira, R. (1998). *Definite description resolution in unrestricted texts*. Unpublished doctoral dissertation, Centre for Cognitive Science, University of Edinburgh.

Vieira, R., & Poesio, M. (2000). An empirically based system for processing definite descriptions. *Computational Linguistics*, *26*(4), 539–593.

Weiner, E. J., & Labov, W. (1983). Constraints on the agentless passive. *Journal of Linguistics*, *19*(1), 29–58.

West, B. T., Welch, K. B., & Galecki, A. T. (2007). *Linear Mixed Models: A practical guide using statistical software*. Boca Raton, FL, USA: Chapman & Hall/CRC.

Wolk, C., Bresnan, J., Rosenbach, A., & Szmrecsányi, B. (2012). Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica*. (To appear.)

Xing, X., Zhang, Y., & Han, M. (2010). Query difficulty prediction for Contextual Image Retrieval. In C. Gurrin et al. (Eds.), *Proceedings of the 32nd European Conference on Information Retrieval (ECIR 2010)* (pp. 581–585).

Appendix: Manual annotation of the features

Animacy of recipient (AnRec)

Following Bresnan et al. (2007), the animacy of the recipient was annotated as a binary feature: it was labelled either *animate* (human and animal) or *inanimate* (not human or animal). Companies and organizations were considered animate when it was evident from the context that the writer meant the people working in these institutions.

Concreteness of theme (ConTh)

For the annotation of the concreteness of the theme, the instructions in Bresnan et al. (2007) were not very clear, except that the feature again allowed only two values: either *concrete* or *abstract*. We decided to follow Garretson (2003), in which a noun phrase is deemed concrete if it is prototypically concrete. We assumed that prototypically concrete objects have a known physical size. The themes that did not fit this description were labelled *abstract*.

Definiteness of recipient and theme (DefRec, DefTh)

For both the recipient and the theme we annotated the definiteness. All (syntactic) object heads that were preceded by a definite article, a genitive form or a definite pronoun (e.g. demonstrative and possessive pronouns), and all objects that were proper nouns or definite pronouns themselves, were annotated as *definite*. The remaining objects were given the value *indefinite*.

Discourse givenness of recipient and theme (GivRec, GivTh)

A recipient or theme was labelled *given* when it was mentioned in the preceding context (maximally 20 clauses before). We also considered an object given when it was stereotypical of something mentioned in the preceding context, or when it was part of the writing context (e.g. the newspaper article itself, or the reader). *You*, *one* and *us* as impersonal pronouns were annotated as given as well. All remaining objects received the value *new*.

Number of recipient and theme (NrRec, NrTh)

Recipients and themes were annotated for number: *singular* or *plural*. In case a recipient or theme could refer to something singular or plural (which is especially the case with the pronoun *you*), the antecedent was checked.

Person of recipient (PrsRec)

Person of recipient was annotated by giving it the value *local* or *nonlocal*. Local recipients are in first or second person (e.g. *I*, *me*, *yourself*), non-local ones in third person.

Pronominality of recipient and theme (PrnRec, PrnTh)

We also annotated whether the recipient and the theme were (syntactically) headed by a pronoun and thus *pronominal*, or not (*nonpronominal*). We treated all types of pronouns as such, including for instance indefinite and relative pronouns like *all* and *that*.

Length difference (LenDif)

An important factor in clause word order is the so-called *principle of end weight* (e.g., Quirk et al., 1972), which states that language users tend to place the more complex constituents at the end of an utterance. Bresnan et al. (2007) therefore included a feature indicating the length difference between the recipient and the theme. Following their approach, we used a Perl script that counts the number of words in the recipient and the theme by splitting on white space, and takes the natural log of these lengths to smoothen outliers. The length difference is then calculated by subtracting the recipient length from the theme length.

Nederlandse samenvatting

Wanneer we taal gebruiken, maken we bewust of onbewust keuzes over de woordvolgorde en de grammaticale structuur van een zin. Kijk bijvoorbeeld naar deze twee zinnen:

- De boze koningin geeft de giftige appel aan Sneeuwwitje.
- De boze koningin geeft Sneeuwwitje de giftige appel.

Deze twee zinnen lijken erg op elkaar. Soms kiest een spreker of schrijver voor de ene vorm en soms voor de andere vorm. Dit proefschrift gaat over deze keuze, die ook wel de datiefalternantie genoemd wordt. De datiefalternantie bestaat niet alleen in het Nederlands, maar bijvoorbeeld ook in het Engels, het Spaans en het Grieks. In dit proefschrift gaat het om de datiefalternantie in het Engels. Uit eerder onderzoek is gebleken dat mensen (onbewust) vaak eerst de makkelijk te verwerken dingen noemen: dingen die al eerder genoemd zijn, die concreet zijn, weinig woorden bevatten, etc. Daarna noemen we pas de dingen die nieuw zijn, die abstract zijn of veel woorden bevatten. Bovendien is vaak gevonden dat het werkwoord van de zin (*geeft* in de voorbeeldzinnen) de keuze beïnvloedt.

Er is al veel onderzoek gedaan naar de datiefalternantie. Vaak wordt hiervoor gebruik gemaakt van *corpora*: databases van taal, bijvoorbeeld uit kranten en boeken, of uit opgenomen gesprekken. Hierin zijn de datiefzinnen opgezocht, waarna is genoteerd wat de eigenschappen zijn van de twee objecten: *de giftige appel* (dit object noemen we het thema) en *Sneeuwwitje* (dit object noemen we de ontvanger). Zo kun je een dataset verzamelen met voorbeelden van datiefzinnen, om hiermee te onderzoeken welke eigenschappen de keuze beïnvloeden.

Het is niet eenduidig *hoe* je dit alles zou moeten onderzoeken. Hoe kies je welke eigenschappen je mee moet nemen in het onderzoek? En hoe definieer je deze eigenschappen, zodat andere onderzoekers het onderzoek kunnen herhalen? Hoe kun je ervoor zorgen dat de dataset groot genoeg is om onderzoek op te kunnen doen? Hoe ga je ermee om dat de datiefzinnen zijn gebruikt door mensen verschillend in leeftijd, geslacht en nationaliteit? Met welke statistische methode onderzoek je de dataset eigenlijk, en hoe interpreteer je de resultaten? Deze vragen heb ik behandeld in vijf afzonderlijke hoofdstukken in dit proefschrift.

De algemene conclusie is dat de keuzes die onderzoekers maken vaak invloed hebben op de resultaten van het onderzoek, en daarmee op de conclusies

die getrokken worden. Dit is geen verrassende uitkomst, maar wel een belangrijke. Als we beter willen begrijpen hoe de datiefalternantie werkt, zullen de verschillende onderzoeken vergelijkbaar moeten zijn. Hetzelfde geldt voor de verzamelde data. Daarom presenteert dit proeschrift ook een manier voor (semi-)automatische dataverzameling, waardoor het eenvoudiger is om grote hoeveelheden consistente data te verkrijgen.

Een andere belangrijke bevinding is dat de keuze van een datiefconstructie grotendeels verklaard kan worden door oppervlakkige informatie zoals het werkwoord van de zin, het aantal woorden in de twee objecten en de eigenschappen van degene die de zin geuit heeft. Om vast te kunnen stellen of de complexere eigenschappen zoals concreetheid van de objecten een rol spelen in dit soort taalkeuzes, zullen we in de toekomst niet alleen moeten kijken naar de Engelse datiefalternantie, maar ook naar andere alternanties en andere talen.

Hieronder volgt per hoofdstuk een korte samenvatting.

Hoofdstuk 2: Selectie van eigenschappen

Dit hoofdstuk geeft een vergelijking van zes manieren om een logististisch-regressie-model te maken voor het onderzoeken van de Engelse datiefalternantie. In totaal worden 29 eigenschappen van de twee objecten meegenomen in het onderzoek, en worden ze op drie verschillende manieren toegevoegd aan het regressiemodel. We maken op deze manier drie modellen waarin het werkwoord van de zin ook opgenomen wordt in het model, en drie modellen waarin we het werkwoord negeren.

De modellen waarin het werkwoord meegenomen wordt voorspellen de gegevens het beste. Wat betreft de selectie van de eigenschappen, zou je verwachten dat dezelfde eigenschappen altijd naar voren zullen komen, ongeacht de manier van selectie. Dit blijkt echter niet het geval: de zes manieren leiden tot zes verschillende modellen. Vermoedelijk komt dit doordat de gebruikte dataset niet zo heel groot is: 930 datiefzinnen. Het is moeilijk om te bepalen hoe je moet kiezen tussen de verschillende selectiemethoden en de verschillende modellen, vooral als je ze wil gebruiken om iets te kunnen zeggen over hoe taal werkt. Echte harde conclusies kunnen dus pas getrokken worden als vergelijkbare modellen zijn gevonden met verschillende datasets.

Hoofstuk 3: Definiëren van de eigenschap concreetheid

Dit hoofdstuk vergelijkt verschillende benaderingen om zelfstandige naamwoorden automatisch (met de computer) in te delen op concreetheid. De concreetheid van een woord is erg afhankelijk van de context: de tafel waarin je

zit te werken is bijvoorbeeld concreter dan de (rekenkundige) tafel van zeven. Het hoofdstuk bestaat uit drie studies: (1) een onderzoek waarin we gebruik maken van 68,848 woorden uit een corpus waarin de woordenboekbetekenissen zijn aangebracht (het SemCor Corpus), (2) een onderzoek naar de invloed van de benadering op een regressiemodel van de datiefalternantie (met de data uit Hoofdstuk 2), en (3) een beoordelingsexperiment waarin we onderzoeken of verschillende mensen het eens zijn over de concreetheid van woorden (via het platform Amazon Mechanichal Turk).

De belangrijkste bevindingen zijn dat er voor zowel de automatische (computer)-benaderingen als voor mensen gevallen zijn die overduidelijk concreet of abstract zijn, maar ook veel twijfelgevallen. Bovendien zijn er verschillen tussen de computerbenaderingen onderling, en heeft de keuze van de benadering ook invloed op de gevonden modellen voor de datiefalternantie. Dit betekent dat het, met het oog op vergelijkbaarheid en herhaalbaarheid, misschien niet verstandig is om een gecompliceerde eigenschap als concreetheid mee te nemen in onderzoek naar de datiefalternantie.

Hoofdstuk 4: Automatisch data verzamelen

Om de datiefalternantie te onderzoeken, is het belangrijk om voldoende data te hebben (zie Hoofdstuk 2). Bovendien is het belangrijk dat het onderzoek herhaalbaar is, en vergelijkbaar met andere onderzoeken (zie Hoofdstuk 3). Om die reden presenteert dit hoofdstuk een manier om (semi-)automatisch data te verzamelen voor onderzoek naar de datiefalternantie. Door automatische procedures is het makkelijk om snel veel data te verzamelen, en is deze data altijd consistent.

Het automatisch verzamelen van datiefdata is opgesplits in twee stappen: (1) het vinden van datiefzinnen, en (2) het vaststellen wat de eigenschappen zijn van de twee objecten in deze zinnen. Voor de eerste stap heb ik gebruik gemaakt van een al bestaand systeem dat zinnen automatisch syntactisch kan ontleden: de Connexor parser. Deze parser blijkt helaas veel fouten te maken, waardoor de zinnen met de hand nagekeken moeten worden. Voor de tweede stap heb ik zelf een Perl script geschreven dat de eigenschappen van de objecten opzoekt. Uit de evaluaties blijkt dat de kwaliteit hiervan goed is, al zijn de complexere eigenschappen (zoals concreetheid) niet helemaal overtuigend. De automatische dataverzameling is toegepast in Hoofdstuk 6.

Hoofdstuk 5: Vergelijken van verschillende groepen mensen

Dit hoofdstuk onderzoekt de invloed van de kenmerken van de taalgebruikers zelf, ofwel van de verschillende groepen mensen, op de keuze tussen de twee

datiefzinnen. De focus ligt hierbij op de invloed van de leeftijd, het geslacht en de nationaliteit van de taalgebruiker. De taalgebruikers in dit onderzoek komen uit het Verenigd Koninkrijk, de Verenigde Staten en Australië. Het hoofdstuk bestaat uit twee studies: een corpusonderzoek en een beoordelingsexperiment.

Het Amerikaans-Engels en het Australisch-Engels zijn op verschillende momenten in de geschiedenis ontstaan uit het Brits-Engels. Ondanks de wereldwijde communicatie in onze moderne maatschappij, vinden we andere verschillen tussen Amerikaans- en Brits-Engels dan tussen Australisch- en Brits-Engels. Het aantal woorden in de objecten is bijvoorbeeld meer van invloed op de datiefalternantie in Australië dan in de twee andere landen. De belangrijkste bevinding m.b.t. leeftijd is dat oudere Britten minder moeite blijken te hebben met zinnen als *give the man it* (*geef de man het*) dan jongere Britten en deelnemers uit de andere twee landen. Aangezien deze constructie vooral voorkomt in dialecten, lijkt het erop dat de jonge Britten meer richting standaard taalgebruik gaan. Een andere belangrijke conclusie is dat het grootste deel van de gemaakte datiefkeuzes verklaard kan worden door de kenmerken van de taalgebruiker en door het werkwoord van de zin. De eigenschappen van de objecten lijken een minimale invloed op de keuze te hebben.

Hoofdstuk 6: Interpreteren van statistische modellen

In dit hoofdstuk hebben we de automatische dataverzameling van Hoofdstuk 4 toegepast op een groot corpus (het BNC), met als resultaat een dataset met 11,784 datiefzinnen. De dataset is gebruikt om de gebruikelijke logistische-regressie-modellen te vergelijken met twee andere modellen: een Bayesiaans netwerk en een geheugengebaseerd-leren-model. Deze twee modellen worden vaak in verband gebracht met de cognitieve processen in onze hersenen. Omdat taalverwerking ook een cognitief proces is, is het aantrekkelijk om deze technieken verder te onderzoeken.

Het Bayesiaanse netwerk lijkt op een stroomschema, waarin de eigenschappen van de objecten opgenomen zijn als knopen in het netwerk. Het model geeft even goede resultaten als een regressiemodel. Een van de voordelen is dat er verborgen knopen toegevoegd kunnen worden die de eigenschappen per object samenvatten. Een nadeel is dat je zelf het netwerk moet vormgeven, waardoor er minder kans is op nieuwe inzichten.

In geheugengebaseerd leren wordt ervan uitgegaan dat mensen geen complexe eigenschappen opslaan, maar alleen de woorden zelf. Een geheugengebaseerd-leren model haalt met alleen woorden resultaten die vergelijkbaar zijn met het regressiemodel dat complexe eigenschappen gebruikt. De benadering blijkt vooral moeite te hebben met meer unieke objecten, misschien

omdat mensen meer opslaan dan alleen de woorden. Het zou echter ook kunnen dat de gebruikte dataset te klein is in vergelijking met de data waaraan mensen blootstaan.

De drie modellen presteerden beter dan een simpel model met alleen het werkwoord en het aantal woorden in de twee objecten. Echter, we zien wederom dat het grootste deel van de datiefalternantie met deze oppervlakkige informatie verklaard kan worden.

Curriculum Vitae

Daphne Theijssen was born in Uden (The Netherlands) on 25 June 1984. She finished *Gymnasium* secondary school in 2002, with focus areas Nature & Health and Nature & Technology. In 2005 she obtained a Bachelor of Arts degree in English Language and Culture, and in 2007 a Master of Arts degree in Language and Communication, both at Radboud University Nijmegen.

From 2008 to 2011, she was a PhD student at the Centre for Language Studies at Radboud University Nijmegen. She was part of the research group's colloquium committee and the faculty's PhD council. She spent three months at the Linguistics Department at Stanford University (California, USA) in the Fall of 2010, for which she received a Fulbright grant.

Despite her interest in research and language, she decided to leave academia and linguistics after finishing her dissertation. Since February 2012, she works at PwC Assurance in Breda (the Netherlands), and she will be studying Accounting from September 2012.